COMPUTER SYSTEM AND SECONDARY STORAGE DEVICE

Patent number: JP7073090

Publication date: 1995-03-17

Inventor: MATSUNAMI NAOTO (JP); ISONO SOICHI (JP); MATSUMOTO JUN (JP); YOSHIDA MINORU (JP)

Applicant: HITACHI LTD (JP)

Classification:

- international: ***G06F12/00;*** **G06F12/00;** (IPC1-7): G06F12/00; G06F12/00

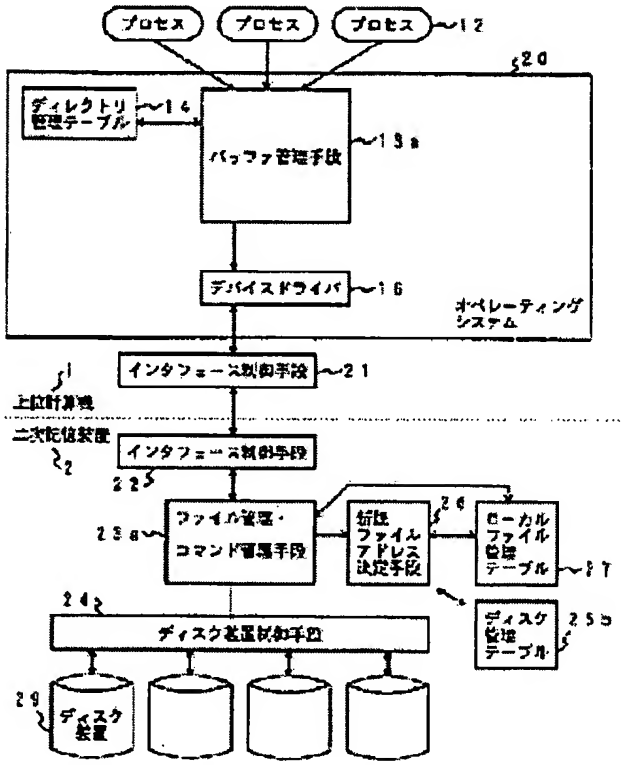
- european:

Application number: JP19940137225 19940620

Priority number(s): JP19940137225 19940620; JP19930149467 19930621

Abstract of **JP7073090**

PURPOSE:To provide a computer system which can arrange a file in an optimum place regardless of the type of a secondary storage device. CONSTITUTION:When an instruction for the new storage of the file is given by a command including identification information of the file from a host computer 1, the file management/command management means 23a of the secondary storage device 2 informs a new file address deciding means 26 of it. The new file address deciding means 26 decides an optimum area as the area for registering the new file by considering the characteristic of the file and the constitution of the self secondary storage device among free areas grasped by referring to a disk management table 25b. The address of the decided area is registered in a local file management table 27 by making it correspond to identification information of the file. The file management/command management means 23a stores the file in the decided area of a disk device 29 through a disk device control means 24.



(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（A）　(11)特許出願公開番号

特開平7－73090

(43)公開日　平成7年(1995)3月17日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　12/00	５２０　Ｊ	8944－5B		
	５０１　Ｈ	8944－5B		

審査請求　未請求　請求項の数14　ＯＬ　（全 36 頁）

(21)出願番号　特願平6－137225

(22)出願日　平成６年(1994)６月20日

(31)優先権主張番号　特願平5－149467

(32)優先日　平５(1993)６月21日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　松並　直人
神奈川県川崎市麻生区王禅寺1099　株式会社日立製作所システム開発研究所内

(72)発明者　磯野　聡一
神奈川県川崎市麻生区王禅寺1099　株式会社日立製作所システム開発研究所内

(72)発明者　松本　純
神奈川県川崎市麻生区王禅寺1099　株式会社日立製作所システム開発研究所内

(74)代理人　弁理士　富田　和子

最終頁に続く

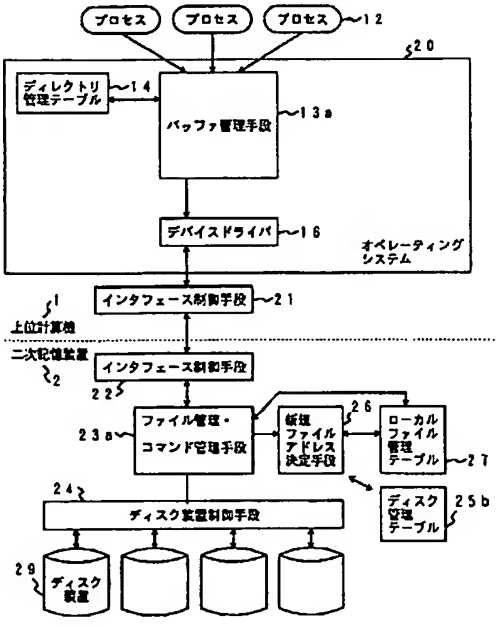
(54)【発明の名称】　計算機システムおよび二次記憶装置

(57)【要約】

【目的】二次記憶装置の種別によらずに、ファイルの最適配置を行うことのできる計算機システムを提供する。

【構成】二次記憶装置２のファイル管理・コマンド管理手段２３ａは、ファイルの識別情報を含んだコマンドによってファイルの新規格納を上位計算機１より指示されると、これを新規ファイルアドレス決定手段２６に通知する。新規ファイルアドレス決定手段２６は、ディスク管理テーブル２５ｂを参照して把握した空き領域のうち、ファイルの特性や自二次記憶装置の構成等を考慮して最適な領域を、新規ファイルを登録する領域として決定する。また、決定した領域のアドレスを、ローカルファイル管理テーブル２７にファイルの識別情報に対応付けて登録する。ファイル管理・コマンド管理手段２３ａは、ディスク装置制御手段２４を介してディスク装置２９の決定された領域にファイルを格納する。

図１１



1

【特許請求の範囲】

【請求項1】計算機と、前記計算機に接続された1または複数の二次記憶装置とを有し、前記二次記憶装置は、二次記憶装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する計算機システムであって、

前記二次記憶装置は、1または複数の論理ブロックより構成されるファイルを記憶する記憶手段と、既に前記記憶手段に記憶されている、前記計算機より要求されたローカルアドレスの論理ブロックのアクセスを実行する手段と、新たな論理ブロックの前記記憶手段への記憶を前記計算機より要求された場合に、記憶を要求された論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段と、決定したローカルアドレスに記憶を要求された論理ブロックを記憶する手段と、決定したローカルアドレスを前記計算機に通知する手段とを有し、

前記計算機は、二次記憶装置に記憶されているファイルを構成する各論理ブロックが記憶されているローカルアドレスを、各論理ブロックに対応付けて管理するファイル管理テーブルと、既に二次記憶装置に記憶されている論理ブロックにアクセスする場合に、前記ファイル管理テーブルを参照し、アクセスする論理ブロックのローカルアドレスを求め、求めたローカルアドレスの論理ブロックのアクセスを二次記憶装置に要求する手段と、新たな論理ブロックの二次記憶装置への記憶を行う場合に、ローカルアドレスを指定せずに、前記二次記憶装置に新たな論理ブロックの記憶を要求する手段と、前記二次記憶装置から通知されたローカルアドレスを、記憶を要求した新たな論理ブロックに対応付けて前記ファイル管理テーブルに登録する手段とを有することを特徴とする計算機システム。

【請求項2】計算機と、前記計算機に接続された1または複数の二次記憶装置とを有し、前記二次記憶装置は、二次記憶装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する計算機システムであって、

前記二次記憶装置は、1または複数の論理ブロックより構成されるファイルを記憶する記憶手段と、前記記憶手段に記憶されているファイルを構成する論理ブロックが記憶されているローカルアドレスを、当該ファイルを指定するファイル識別子と論理ブロックに対応付けて管理するローカルファイル管理テーブルと、記憶を要求されたデータが前記記憶手段に既に記憶されている論理ブロックに属する場合に、ファイルとファイル内の相対アドレスとデータ長とを指定して前記計算機より、アクセスを要求されたデータのローカルアドレスを、指定されたファイル識別子と相対アドレスとデータ長と、前記ファイル管理テーブルより求める手段と、求めたローカルアドレスに記憶されているデータにアクセスする手段と、前記記憶手段へのデータの記憶を前記計算機より要求された場合であって記憶を要求されたデータが前記記憶手段に既に記憶されている論理ブロックに属さないものである場合に、ファイル識別子とファイル内の相対アドレスとデータ長とを指定して前記計算機よりアクセスを要求されたデータが属する新たな論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段と、決定したローカルアドレスに新たな論理ブロックを記憶する手段と、新規のファイルのデータ前記記憶手段への記憶を要求された場合に、当該新規なファイルを構成する論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段と、決定したローカルアドレスに登録を要求された新規なファイルを構成する論理ブロックを記憶する手段と、決定されたローカルアドレスを、当該ローカルアドレスに記憶した論理ブロックと、当該論理ブロックが属するファイル識別子に対応付けて、前記ローカルファイル管理テーブルに登録する手段とを有し、

前記計算機は、既に二次記憶装置に記憶されているファイルのデータにアクセスする場合に、アクセスするデータの属するファイルに割り当てたファイル識別子とアクセスするデータのファイル内の相対アドレスとアクセスするデータ長とを指定して二次記憶装置に当該データのアクセスを要求する手段と、新規なファイルのデータの二次記憶装置への記憶を行う場合に、当該ファイルに新たなファイル識別子を割り当て、割り当てたファイル識別子と記憶するデータのファイル内の相対アドレスと記憶するデータ長とを指定して、前記二次記憶装置に当該新規なファイルの記憶を要求する手段とを有することを特徴とする計算機システム。

【請求項3】データを記憶する二次記憶装置であって、二次記憶装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する二次記憶装置を1または複数接続した計算機であって、

前記計算機は、前記二次記憶装置に記憶されているファイルを構成する各論理ブロックが記憶されているローカルアドレスを、各論理ブロックに対応付けて管理するファイル管理テーブルと、接続している前記二次記憶装置が、論理ブロックを記憶するローカルアドレスを決定する能力を有しているか否かを判別する手段と、二次記憶装置に既に記憶されている論理ブロックにアクセスする場合に、前記ファイル管理テーブルを参照して得たアクセスする論理ブロックのローカルアドレスを指定して二次記憶装置に当該論理ブロックのアクセスを要求する手段と、新たな論理ブロックの、論理ブロックを記憶するローカルアドレスを決定する能力を有している二次記憶装置への記憶を行う場合に、ローカルアドレスを指定せずに、前記二次記憶装置に新たな論理ブロックの記憶を要求する手段と、新たな論理ブロックの記録の要求に対して、前記論理ブロックを記憶するローカルアドレスを決定する能力を有している二次記憶装置から通知された

3

ローカルアドレスを、記憶を要求した新たな論理ブロックに対応付けて前記ファイル管理テーブルに登録する手段と、新たな論理ブロックの、論理ブロックを記憶するローカルアドレスを決定する能力を有していない二次記憶装置への記憶を行う場合に、当該新たな論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段と、決定したローカルアドレスを前記ファイル管理テーブルに記憶する手段と、決定したローカルアドレスを指定して、論理ブロックを記憶するローカルアドレスを決定する能力を有していない二次記憶装置へ当該新たな論理ブロックの記録を要求する手段とを有することを特徴とする計算機。

【請求項4】データを記憶する二次記憶装置であって、二次記憶装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する二次記憶装置を1または複数接続した計算機であって、

前記計算機は、前記二次記憶装置に記憶されているファイルを構成する各論理ブロックが記憶されているローカルアドレスを、各論理ブロックに対応付けて管理するファイル管理テーブルと、接続している前記二次記憶装置が、論理ブロックを記憶するローカルアドレスを決定する能力を有しているか否かを判別する手段と、二次記憶装置に既に記憶されている論理ブロックにアクセスする場合に、前記ファイル管理テーブルを参照して得たアクセスするデータのローカルアドレスを指定して二次記憶装置に当該論理ブロックのアクセスを要求する手段と、新たな論理ブロックの、論理ブロックを記憶するローカルアドレスを決定する能力を有していない二次記憶装置への記憶を行う場合に、当該新たな論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段と、決定した論理ブロックのローカルアドレスを前記ファイル管理テーブルに当該論理ブロックと対応付けて記憶する手段と、決定したローカルアドレスを指定して、論理ブロックを記憶するローカルアドレスを決定する能力を有していない二次記憶装置へ当該新たな論理ブロックの記録を要求する手段と、新規なファイルの論理ブロックを記憶するローカルアドレスを決定する能力を有している二次記憶装置への記憶を行う場合に、ファイルを表すファイル識別子とアクセスするデータのファイル内の相対アドレスとアクセスするデータ長とを指定して、前記二次記憶装置に当該ファイルの記憶を要求する手段とを有することを特徴とする計算機。

【請求項5】計算機に接続され、自装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する二次記憶装置であって、

1または複数の論理ブロックより構成されるファイルを記憶する記憶手段と、新たな論理ブロックの前記記憶手段への記憶を前記計算機より要求された場合に、記憶を要求された論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段と、決定したローカルアドレスに記憶を要求された論理ブロックを記憶する手段とを有することを特徴とする二次記憶装置。

4

【請求項6】計算機に接続され、自装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する二次記憶装置であって、

1または複数の論理ブロックより構成されるファイルを記憶する記憶手段と、前記記憶手段に記憶されているファイルを構成する論理ブロックが記憶されているローカルアドレスを、当該ファイルと各論理ブロックに対応付けて管理するローカルファイル管理テーブルと、前記計算機より前記記憶手段に既に記憶されている論理ブロックに属さないデータの記録を要求された場合に、当該データが属する新たな論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段と、決定したローカルアドレスに新たな論理ブロックを記憶する手段と、決定されたローカルアドレスを、当該ローカルアドレスに記憶した論理ブロックと、当該論理ブロックが属するファイルとに対応付けて、前記ローカルファイル管理テーブルに登録する手段とを有することを特徴とする二次記憶装置。

【請求項7】請求項1、2、3または4記載の計算機システムであって、

前記二次記憶装置の記憶手段は、ディスクアレイ装置であって、

前記論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段は、論理ブロックに属するデータのアクセスを、より並列に行うことができるようなローカルアドレスを所定の手順により求める手段であることを特徴とする計算機システム。

【請求項8】請求項1、2、3または4記載の計算機システムであって、

前記二次記憶装置の記憶手段は、複数の種別の異なる複数の記憶装置であって、

前記論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段は、記録する新たな論理ブロックの属するファイルの特性と、前記複数の記憶装置の特性に応じて、新たな論理ブロックのローカルアドレスを決定することを特徴とする計算機システム。

【請求項9】計算機と、前記計算機に接続された1または複数の二次記憶装置とを有し、前記二次記憶装置は、二次記憶装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する計算機システムにおいて、ファイルをアクセスする方法であって、

二次記憶装置に、ファイルを識別するためのファイル識別情報とファイルを記憶したローカルアドレスを対応付けて記憶するローカルファイル管理テーブルを備えるステップと、

計算機が、二次記憶装置に、新たなファイルの記録を、ファイルの識別情報を指定して要求するステップと、

二次記憶装置が、記録を要求された新たなファイルを記

5

憶するローカルアドレスを決定するステップと、
二次記憶装置が、前記記憶媒体の決定したローカルアドレスに新たなファイルを記憶するステップと、
二次記憶装置が、決定したローカルアドレスを指定されたファイルの識別情報に対応付けて、前記ローカルファイル管理テーブルに記憶するステップとを有することを特徴とするファイルのアクセス方法。

【請求項１０】請求項９記載のファイルのアクセス方法であって、
二次記憶装置が、決定したローカルアドレスを前記計算機に通知するステップと、
計算機が通知されたローカルアドレスを、記録を要求したファイルに対応付けて管理するステップと、
計算機に、ファイルを識別するためのファイル識別情報とファイルが記憶されているローカルアドレスを対応付けて記憶するファイル管理テーブルを備えるステップと、
計算機が、既に二次記憶装置に記憶されているファイルのアクセスを行う場合に、前記ファイル管理テーブルより、当該ファイルに対応付けて管理しているローカルアドレスを求めるステップと、
計算機が、既に二次記憶装置に記憶されているファイルのアクセスを、求めたローカルアドレスを指定して二次記憶装置に要求するステップと、
二次記憶装置が、既に二次記憶装置に記憶されているファイルのアクセスを要求された場合に、要求で指定されたローカルアドレスのファイルにアクセスするステップとを有することを特徴とするファイルのアクセス方法。

【請求項１１】計算機と、前記計算機に接続された１または複数の二次記憶装置とを有し、前記二次記憶装置は、二次記憶装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する計算機システムであって、
前記計算機は、前記二次記憶装置から通知されたローカルアドレスを、前記二期記憶装置の前記通知されたローカルアドレスに記憶したデータの論理ブロックに対応付けして記録するファイル管理テーブルを備えたことを特徴とする計算機システム。

【請求項１２】計算機と、前記計算機に接続された１または複数の二次記憶装置とを有し、前記二次記憶装置は、二次記憶装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する計算機システムにおいて、前記計算機が前記二次記憶装置にファイルを記憶する方法であって、
前記計算機は、ファイルを構成する論理ブロックを前記二次記憶装置に新規に記憶する際に、二次記憶装置において前記論理ブロックを記憶するローカルアドレスを決定することを要求する識別子と、前記ファイルの識別子と、論理ブロックの論理アドレスとを含む論理ブロックの格納要求を前記二次記憶装置に発行し、

6

前記二次記憶装置は、前記論理ブロックの格納要求に応じて決定したローカルアドレスを、前記論理ブロックの格納要求に応答して、前記計算機に発行することを特徴とする方法。

【請求項１３】計算機と、前記計算機に接続された１または複数の二次記憶装置とを有し、前記二次記憶装置は、二次記憶装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する計算機システムにおいて、前記計算機が前記二次記憶装置のデータにアクセスする方法であって、
前記計算機は、前記二次記憶装置に記憶されている、ファイルのデータにアクセスするする際に、前記ファイルの識別子と、前記データの前記ファイルの先頭よりの相対アドレスと、要求するデータのサイズとを含むアクセス要求を発行し、
前記二次記憶装置は、前記アクセス要求に含まれる記ファイルの識別子と、前記データの前記ファイルの先頭よりの相対アドレスと、要求するデータのサイズに対応するローカルアドレスのデータにアクセスし、前記アクセス要求に対する結果の応答として、終了ステータスを前記計算機に発行することを特徴とする方法。

【請求項１４】計算機に接続され、自装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する二次記憶装置であって、
１または複数の論理ブロックより構成されるファイルを記憶する記憶手段と、新たな論理ブロックの前記記憶手段への記憶を前記計算機より要求された場合に、記憶を要求された論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段と、決定したローカルアドレスに前記計算機に通知する手段とを有することを特徴とする二次記憶装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、上位計算機と二次記憶装置とから構成する計算機システムに関し、特に、二次記憶装置におけるファイルの最適配置の技術に関するものである。

【０００２】

【従来の技術】上位計算機と二次記憶装置とから構成される計算機システムとしては、ＵＮＩＸとして知られるオペレーティングシステム（以下、「ＯＳ」と記す）を採用するワークステーション（以下、「ＷＳ」と記す）にハードディスク装置（以下、「ディスク装置」という）を接続したシステムが知られている。

【０００３】また、このような計算機システムにおけるファイル管理の技術としては、「ＵＮＩＸ４．３ＢＳＤの設計と実装」（丸善、１９９１年出版）に記載された技術が知られている。

【０００４】以下、この技術について説明する。

【０００５】図２４に、この計算機システムの構成を示

す。

【0006】図中、1が上位計算機、2が二次記憶装置である。

【0007】上位計算機1において、20はOS，12はOS20が管理するユーザが実行したアプリケーションプログラム（以下、「AP」と記す）のプロセスを表している。すなわち、上位計算機において実行されているアプリケーションプログラムはOS20によりプロセス12として管理されている。

【0008】また、OS20において、13はファイルの管理やファイルアクセス時のデータ転送に使用するバッファを管理するファイル管理・バッファ管理手段、15はディレクトリやファイル論理的なアドレス（OS内部でのアドレス）と、ディスク装置内でのローカルなアドレス（ローカルアドレス）との対応付けのための情報を記述するファイル管理テーブル、14はディレクトリ情報（ディレクトリやファイルの名称とファイル管理テーブルの対応付けのための情報）を記述するディレクトリ管理テーブルを示している。また、16はファイル管理・バッファ管理手段13からのファイルアクセス要求を、様々な二次記憶装置等の外部デバイスの物理特性に合わせて変換、制御を行うデバイスドライバ、30はファイルを新規にライトする際に書き込みに最適なディスク装置のローカルアドレスを決定するファイルアドレス決定手段、25aはディスクの使用状況を管理するディスク管理テーブル、21、22は上位計算機と二次記憶装置との間の通信およびデータ転送を制御するインタフェース制御手段を示している。

【0009】ここで、論理アドレスとは、上位計算機においてファイルを管理するための論理的なアドレスである。また、ローカルアドレス（ディスク装置内でのローカルなアドレス）とは、ディスク装置（二次記憶装置）が、データを管理するためのディスク装置内でのローカルなアドレスであり、論理的なアドレスである場合も、物理的なアドレスである場合もある。たとえば、SCSI (Snall Computer Interface)に適合したディスク装置（SCSIディスク）では、ローカルアドレスとして論理的な連続番号（LBA；Logical Block Address)を用いている。また、IDE (Intelligent Device Electronics)ディスク装置等では、ローカルアドレスとして、ヘッド番号、シリンダ番号、セクタ番号等の、物理的な記憶位置を直接表す物理アドレスを用いている。

【0010】次に、二次記憶装置2において、29は1台以上のディスク装置、24はディスク装置29を制御するディスク装置制御手段、23は上位計算機から送信されたリード及びライトコマンドを受信、解析し、ディスク装置29特有のディスクコマンドを作成してディスク装置制御手段24に送るコマンド管理手段を示している。

【0011】以下、プロセス12が既存のファイルにアクセスを行う際に実行するファイルアクセス処理を説明する。

【0012】プロセス12がファイルアクセスを行う際にはOS20にファイルアクセスリードおよびライト要求（システムコール）を発行する。

【0013】OS20のファイル管理・バッファ管理手段13はこの要求を受け、ディレクトリ管理テーブル14を参照し、ファイル名からファイル管理テーブルの位置を求める。次に求めたファイル管理テーブル15を参照し、プロセスのリード（またはライト）要求データが存在する（またはデータを登録する）論理アドレスを算出し、ローカルアドレスに変換する。また、データ転送に必要なバッファエリアを確保する。

【0014】次に、OS20は、デバイスドライバ16に対し算出したローカルアドレスを用いアクセス依頼を行う。デバイスドライバ16はこのアクセス依頼を受け、ファイルの格納されている（または、ファイルを格納する）二次記憶装置2の物理特性に合わせた形のコマンドを作成し、インタフェース制御手段21を介し、二次記憶装置2に送出する。

【0015】二次記憶装置2ではコマンド管理手段23がインタフェース制御手段22を介しコマンドを受信し、このコマンドを解析する。そして、アクセスの種類、アクセス開始アドレス、データ長等のコマンドの解析結果に基づき、ディスク装置29特有のディスクコマンドを作成し、ディスク制御装置24にコマンド実行を依頼する。ディスク制御装置24はこのコマンドに基づきディスク装置29を制御し、適切なデータ転送処理を実行する。

【0016】さて、プロセス12が既存のファイルにアクセスを行う際には二次記憶装置2の、どのアドレスにファイルを格納するかを決定する必要がある。

【0017】そこで、新規にファイルを作成する際には、以下に示す新規登録処理を、ファイルアクセス（ライト）処理に先立ち行う必要がある。

【0018】すなわち、新規ファイルのライト要求をプロセス12から依頼されたとき、ファイル管理・バッファ管理手段13は、ディレクトリ管理テーブル14にファイル名を登録し、ファイル管理テーブル15にこのファイルのための管理データ領域を確保し、その位置をディレクトリ管理テーブル14に登録する。次に、ファイルアドレス決定手段30は、ディスク管理テーブル25aを参照しながら、予めユーザが設定しておいたディスク装置のヘッド数、シリンダ数、セクタ数等の物理特性パラメータを基に最適アドレス、すなわちできるだけシークを行わず、また無駄な回転待ちの発生しないような領域のローカルアドレスを所定の配置アルゴリズムにより決定する。そして、この決定したローカルアドレスをファイル管理テーブルに登録し、論理アドレスとの対応付けを行う。

9

【0019】以上の新規登録処理により、以降のアクセスにおいては、ディレクトリ管理テーブル14、ファイル管理テーブル15の両者を参照して行う通常のファイルアクセス処理により、このファイルアクセスが可能となる。

【0020】

【発明が解決しようとする課題】前記「UNIX4．3BSDの設計と実装」（丸善、1991年出版）に記載の技術によれば、前述したように、新規に作成したファイルを格納する最適アドレスを決定するために、ディスク装置等の二次記憶装置の物理パラメータを予め設定しておかなければならない。

【0021】一方、近年、大容量化、高性能化、高信頼化の要請に伴い、単に1台または複数台のディスク装置を接続するのではなく、ディスクをアレイ上に配置し、各ディスク装置にデータを分割配置するディスクアレイ装置等、高度化、複雑化した様々な形態の二次記憶装置が出現してきている。

【0022】このため、二次記憶装置の物理パラメータも、ディスク単体のパラメータのみならず、アレイの構成、データ分配方式、高信頼化を実現するために複数台のディスク装置のデータから計算した冗長データの配置等、その二次記憶装置のアーキテクチャに依存し多種多様となる。また、一般的に、上位計算機と二次記憶装置は、独立して流通しているため、製造元であらかじめ二次記憶装置のパラメータを予め上位計算機に設定することはできない。

【0023】このため、二次記憶装置のパラメータの設定はユーザが行うことになるが、ユーザが、これら多種多彩な二次記憶装置のすべてのパラメータを上位計算機に設定することは困難であり、現実には、これを行うことができない場合が多い。その結果、上位計算機において複雑な構成をとる二次記憶装置へのファイル最適配置は困難になり、二次記憶装置の性能を有効に利用することができないという問題が生じる。

【0024】そこで、本発明は、上位計算機に二次記憶装置のパラメータを設定すること無しに、ファイルの二次記憶装置への最適配置を実現することのできる計算機システムを提供することを目的とする。

【0025】

【課題を解決するための手段】前記目的達成のために、本発明は、たとえば、計算機と、前記計算機に接続された1または複数の二次記憶装置とを有し、前記二次記憶装置は、二次記憶装置内のローカルなアドレスであるローカルアドレスによって記憶したデータを管理する計算機システムであって、前記二次記憶装置は、1または複数の論理ブロックより構成されるファイルを記憶する記憶手段と、ローカルアドレスを指定して前記計算機より要求された、既に前記記憶手段に記憶されている論理ブロックのアクセスを実行する手段と、新たな論理ブロックの前記記憶手段への記憶を前記計算機より要求された場合に、記憶を要求された論理ブロックを記憶するローカルアドレスを所定の手順により決定する手段と、決定したローカルアドレスに記憶を要求された論理ブロックを記憶する手段と、決定したローカルアドレスを前記計算機に通知する手段とを有し、前記計算機は、二次記憶装置に記憶されているファイルを構成する各論理ブロックが記憶されているローカルアドレスを、各論理ブロックに対応付けて管理するファイル管理テーブルと、既に二次記憶装置に記憶されている論理ブロックにアクセスする場合に、前記ファイル管理テーブルを参照してアクセスする論理ブロックのローカルアドレスを求め、ローカルアドレスを指定して二次記憶装置に当該論理ブロックのアクセスを要求する手段と、新たな論理ブロックの二次記憶装置への記憶を行う場合に、ローカルアドレスを指定せずに、前記二次記憶装置に論理ブロックの記憶を要求する手段と、前記二次記憶装置から通知されたローカルアドレスを、記憶を要求した新たな論理ブロックに対応付けて前記ファイル管理テーブルに登録する手段とを有することを特徴とする計算機システムを提供する。

【0026】

【作用】本発明に係る計算機システムによれば、計算機は、新たな論理ブロックの二次記憶装置への記憶を行う場合には、二次記憶装置のローカルアドレスを指定せずに、前記二次記憶装置に論理ブロックの記憶を要求する。一方、二次記憶装置は、新たな論理ブロックの前記記憶手段への記憶を前記計算機より要求されると、記憶を要求された論理ブロックを記憶するローカルアドレスを所定の手順により決定し、決定したローカルアドレスに記憶を要求された論理ブロックを記憶すると共に、決定したローカルアドレスを前記計算機に通知する。計算機は、通知されたローカルアドレスを、記憶を要求した新たな論理ブロックに対応付けて前記ファイル管理テーブルに登録し、以降、この論理ブロックにアクセスする場合には、ファイル管理テーブルを参照し、アクセスする論理ブロックのローカルアドレスを指定して二次記憶装置にアクセスを要求する。

【0027】このように、本発明によれば、ファイルを構成する論理ブロックを記憶するローカルアドレスを、二次記憶装置が決定する。ここで、個別の二次記憶装置に、自身の構成や状態を把握させるのは容易であるので、二次記憶装置は、論理ブロックの最適な記憶位置を決定することができる。よって上位計算機に二次記憶装置のパラメータを設定すること無しに、ファイルの二次記憶装置への最適配置を実現することができる。

【0028】

【実施例】以下、本発明に係る計算機システムの実施例を説明する。

【0029】はじめに、本実施例に係る計算機システム

のハードウェア構成例を図１に示す。

【００３０】図示するように、本実施例に係る計算機システムは、上位計算機１と二次記憶装置２から成る。上位計算機１は、ＯＳやアプリケーションプログラムを実行するＣＰＵ１０と、ＣＰＵ１０が用いる一次記憶装置１１と、二次記憶装置２へのインプット／アウトプット要求（Ｉ／Ｏ要求）を制御するＩ／Ｏ制御部９と、二次記憶装置２へのインタフェース８から構成される。

【００３１】一方、二次記憶装置２は、上位計算機１からのＩ／Ｏ要求の授受を行うインタフェース７と、データ転送および二次記憶装置２内部の制御を司る制御装置５と、データを格納するディスク装置４と、ディスク装置４と上位計算機１とのデータ転送速度の違いを吸収するためのバッファ装置とから構成される。

【００３２】ただし、本実施例係る計算機システムのハードウェア構成は、この他、図１に示した構成に準じる他の構成としてもよい。

【００３３】次に、本実施例に係る計算機システムにおける、ファイルの論理的位置と、ファイルの、二次記憶装置上の位置の管理の方式について説明する。

【００３４】図２に示すように、ファイルの論理的位置は、ディレクトリを用いて構築した木構造によって階層的に管理される。これは、前述したＵＮＩＸなどのＯＳで採用している方式である。

【００３５】ここで、ファイルは、二次記憶装置の実記憶領域に、所定長のデータの集まりであるブロック毎に記憶される。そして、ファイルの二次記憶装置上のローカルな位置は、ファイルを構成するブロックの、それぞれが記憶されている論理アドレスと、この論理アドレスに対応する二次記憶装置上のローカルアドレスを管理することにより管理される。二次記憶装置上のローカルな位置は、二次記憶装置によって、二次記憶装置のローカルなアドレスを用いて管理される。二次記憶装置のローカルなアドレスをローカルアドレスと呼ぶ。前述したようにローカルアドレスは二次記憶装置上の物理位置を指定する物理アドレスである場合もあり、二次記憶装置上の論理的な位置を指定する論理的なアドレスである場合もある。

【００３６】但し、このような管理方式に準じた方式によって、ファイルの論理的位置と、ファイルの二次記憶装置上の位置を管理するようにしてもよい。

【００３７】さて、新規ファイルを二次記憶装置２に格納する際には、前述した新規登録処理の一環として、そのファイルを分割したブロックのそれぞれを最適化アルゴリズムに従い２次記憶装置のローカルアドレスにマッピングする処理を行う必要がある。しかし、二次記憶装置２内にディスク装置を複数台接続したり、二次記憶装置２が、いわゆるＲＡＩＤ（Redundant Arrays of Inexpensive Disks)アーキテクチャに従い複数のディスク装置にデータを分散配置するような複雑な構成を持つ二次記憶装置であった場合には、上位計算機１に、使用する二次記憶装置の種別に応じた最適化アルゴリズムや、多種多彩なパラメータを入力することは極めて困難である。このため、通常は、二次記憶装置２を従来の単純な二次記憶装置に見せかけ、従来のアルゴリズムを使用することになる。しかし、このようにすると、二次記憶装置の持つ潜在的な能力を引き出せないばかりでなく最悪の場合従来に比べても低い性能しか期待できない場合がある。

【００３８】以下、本発明の第１の実施例について説明する。

【００３９】図３に、本第１実施例に係る計算機システムの構成を示す。

【００４０】図示するように、本第１実施例に係る計算機システムは、上位計算機１と二次記憶装置２を有している。

【００４１】上位計算機１において、２０はＯＳ、１２はＯＳ２０が管理するアプリケーションプログラムのプロセスを表している。上位計算機において実行されているアプリケーションプログラムはＯＳ２０によりプロセス１２として管理されている。

【００４２】また、ＯＳ２０において、１３はファイルの管理やファイルアクセス時のデータ転送に使用するバッファを管理するファイル管理・バッファ管理手段、１５はディレクトリやファイルの論理アドレスとローカルアドレスとの対応を記述するファイル管理テーブル、１４はディレクトリ情報を記述するディレクトリ管理テーブルを示している。また、１６はファイル管理・バッファ管理手段１３からのファイルアクセス要求を、様々な二次記憶装置等の外部デバイスの物理特性に合わせて変換、制御を行うデバイスドライバ、１８はファイルがオープンされたことを二次記憶装置２に通知するオープンファイル情報通知手段、１９は新規のファイルをライトすることを二次記憶装置２に通知する新規ファイルライト通知手段、１７は二次記憶装置２により決定された新規ファイルの格納アドレスを受信しファイル管理テーブル１５に登録する新規ファイルアドレス登録手段を示している、また、上位計算機１において、２１は二次記憶装置との間の通信およびデータ転送を制御するインタフェース制御手段を示している。

【００４３】ファイル管理テーブル１５はファイル毎に設けられたファイル管理情報を格納している。この、ファイル毎に設けらたファイル管理情報は、ファイル識別番号により特定されるファイル管理情報のアドレスを介して特定される。なお、前記ＵＮＩＸでは、このファイル管理情報をｉノードと呼び、ファイル識別番号をｉノード番号と呼んでいる。

【００４４】各ファイル管理情報は、図４に示すように構成されている。

【００４５】すなわち、各ファイル管理情報は、モード

13

151、所有者152、タイムスタンプ153、大きさ154、ブロック数155、参照カウント156、参照ポインタ157、更新フラグ158のエントリを有している。

【0046】モード151は、ファイルの種類と、そのファイルの現在のアクセスモード（リード／ライト等）の情報を格納し、所有者152には、ファイルの所有者や、そのファイルにアクセス権を持つユーザグループを識別する情報を格納する。また、タイムスタンプ153は、ファイルが最後のアクセスされた時間や、自ファイル管理テーブルが更新された時間等を格納する。大きさ154は、ファイルのバイト数を格納し、ブロック数154はファイルの前記ブロック数を格納する。また、参照カウンタ156は、ファイルが参照された回数を格納し、参照ポインタ177は、ファイルの各ブロックの論理アドレスと二次記憶装置のローカルアドレスとを対応付けて格納する。以下では、ブロックの論理アドレスを論理ブロックアドレス、ブロックの二次記憶装置のローカルアドレスをローカルブロックアドレスという。

【0047】次に、ディレクトリ管理テーブル14も、ディレクトリ毎に設けられたディレクトリ管理情報を格納しており、各ディレクトリ管理情報は、ファイル管理テーブル15の対応するディレクトリに属するファイルに対応するファイル管理情報を特定するファイル識別番号と、対応するディレクトリに属するディレクトリに対応するディレクトリ管理情報を特定する番号の情報（またはディレクトリ管理テーブルの存在位置を示すポインタ）を記憶する。

【0048】このような、ディレクトリ管理テーブル14とファイル管理テーブル15の構成により、デレクトリdir1、dir2、dir3の指定とfile5の指定によりfile5というファイルを次のようにしてfile5を構成する各ブロックのローカルブロックアドレスを得ることができる。すなわち、dir1、dir2、dir3のディレクトリ管理テーブル14を順次走査し、dir3のディレクトリ管理情報から、ファイル管理テーブル15のfile5に対応するファイル管理情報のファイル識別番号を得、file5のファイル管理情報より、file5を構成する各ブロックのローカルブロックアドレスを得る。

【0049】さて、一方、二次記憶装置2において、29は1台以上のディスク装置、24はディスク装置29を制御するディスク装置制御手段、23は上位計算機から送信されたのリード及びライトコマンドを受信、解析し、ディスク装置29特有のディスクコマンドを作成してディスク装置制御手段24に送るコマンド管理手段、22は上位計算機1との間の通信およびデータ転送を制御するインタフェース制御手段、27は二次記憶装置2のディスク装置29に記憶されているファイルのファイル管理情報を格納するローカルファイル管理テーブル、

14

39は上位計算機1の送出するファイルがオープンされたことを示す情報を受信しローカルファイル管理テーブル27に登録するオープンファイル情報受信登録手段、26は上記上位計算機1の送出する新規のファイルをライトすることを受信し新規ファイルの格納アドレスを決定する新規ファイルアドレス決定手段、28は新規ファイルアドレス決定手段26により決定されたアドレスを上記上位計算機1に通知する新規ファイルアドレス通知手段、25bは二次記憶装置内部のディスク装置の構成を決定するパラメータおよびディスク装置の利用状況を管理するディスク管理テーブルである。

【0050】ローカルファイル管理テーブル27の構成およびローカルファイル管理テーブル27に格納されるファイル管理情報は、上位計算機1のファイル管理テーブル15およびファイル管理情報と同じであり、同じファイルのファイル管理情報は同じファイル識別番号で特定される。

【0051】以下、本第1実施例に係る計算機システムのファイルのアクセス動作を説明する。

【0052】まず、あらかじめ、ファイル管理バッファ管理手段13は、二次記憶装置2のローカルファイル管理テーブル27に格納されているファイル管理情報を、上位計算機1のファイル管理テーブル15にロードしておく。このロードは、上位計算機のイニシャル時や、二次記憶装置に新たな記憶媒体がマウントされた時や、次に述べるファイルオープン処理の開始時等に行う。

【0053】さて、APのプロセス12から、オープンシステムコールを発行されると、OS20は、図5に示すファイルオープン処理を行う。

【0054】すなわち、図示するように、オープンシステムコールを発行されると（ステップ101）、OS20内部のファイル管理・バッファ管理手段13は、このシステムコールを解析し、ファイルのアクセスモードを判定する（（ステップ102）。もしライトモードでファイルのオープンを行うのであれば、新規のファイルであるかどうかをさらに判定し（ステップ103）、もし、新規ファイルであるならばディレクトリ管理テーブル14の、このファイルを作成するディレクトリに対応するディレクトリ管理情報にファイル名を登録し、ファイル管理テーブル15中に、新規ファイルについてのファイル管理情報を格納する領域を確保する（ステップ104）。また、この確保したファイル管理情報のファイル識別番号を、ディレクトリ管理テーブル14の、ファイルを作成するディレクトリに対応するディレクトリ管理情報に登録する。そして、ファイル識別番号をプロセス12に通知する（ステップ105）。

【0055】新規のファイルでなければ、ディレクトリ管理テーブル14より、ファイルのファイル識別番号を獲得し、これをプロセス12に通知する（ステップ105）。

15

【００５６】次に、ファイル管理・バッファ管理手段１３はこのファイル固有のファイル識別番号により特定されるファイル管理情報のローカルアドレス、ファイル名等の情報をオープンファイル情報通知手段１８に伝える。オープンファイル情報通知手段１８は、これらの情報をデバイスドライバ１６、インタフェース制御手段２１、２２を介して二次記憶装置２へ通知する（ステップ１０６）。

【００５７】二次記憶装置２のオープンファイル情報受信登録手段３９は、これらの情報を受信し、もし、新規なファイルについてのものである場合にはローカルファイル管理テーブル２７に新たなファイル管理情報の領域を確保し、受信した情報を登録する。

【００５８】以上の処理により、オープンしたファイルの管理情報は上位計算機１のファイル管理テーブル１５と二次記憶装置２のローカルファイル管理テーブル２７両方に登録されていることになる。

【００５９】次に、ファイルのオープン処理が終了すると、ファイルのリードもしくはライト処理を実行する。

【００６０】図６にＯＳ２０におけるファイルのリード処理の手順を、図８にＯＳ２０におけるファイルのライト処理の手順を、図９に二次記憶装置におけるファイルのリード処理とライト処理の手順を示す。

【００６１】まず、ファイルのリード処理について説明する。

【００６２】図６に示すように、上位計算機１において、ＡＰ（プロセス１２）からファイルリードシステムコール（ｒｅａｄ（））が発行されると（ステップ２０１）、ＯＳ２０のファイル管理・バッファ管理手段１３は、ディレクトリ管理テーブル１４を参照し当該ファイルのファイル識別番号（ファイル管理テーブル１５内のファイル管理情報の位置を示す）を獲得する（ステップ２０２）。次に、ファイル管理・バッファ管理手段１３は獲得したファイル識別番号により指定されるファイル管理テーブル１５内のファイル管理情報の情報を得る。

【００６３】さて、ここで、ＡＰは、ファイルリードシステムコールもしくはファイルライトシステムコールによってオフセットと呼ぶバイト単位でファイルのリードやライトを要求する。図７は、ＡＰが、ファイルｆｄについて設けられているファイルポインタｆｐからｎバイトのオフセットをユーザバッファｂｕｆｆｅｒにリードすることを要求している場合を示している。図示するように、このオフセットは論理ブロックアドレス１、２にまたがっているデータであり、すなわち論理ブロック１、２、に対応する２つのローカルブロック（ローカルブロックアドレス＃１１０２０と、＃１１０２８）とからリードする必要がある。ローカルブロックは、論理ブロックに対応する二次記憶装置２上のブロックである。

【００６４】そこで、次に、ファイル管理・バッファ管理手段１３は、得たファイル管理情報より要求されたオフセットのデータが存在する論理ブロックアドレス１、２を求める（ステップ２０３）。そして、これら２つの当該論理ブロックをリードするために必要な一時作業領域としてシステムバッファを確保する（ステップ２０４）。次に、はじめの論理ブロックアドレス１を選択し（ステップ２０５）、これをローカルブロックアドレス＃１１０２０に変換し（ステップ２０６）、デバイスドライバ１６に、このローカルブロックアドレスのデータブロックの転送依頼を発行する。デバイスドライバ１６はこの要求を受け、二次記憶装置固有のリードコマンドを生成し、インタフェース制御手段２１を介し二次記憶装置２に送出する（ステップ２０７）。

16

【００６５】リードコマンドは、二次記憶装置２のコマンド管理手段２３は上記リードコマンドをインタフェース制御手段２２を介して受信される。コマンドを受信した二次記憶装置２は、図９に示すように、まず、このコマンドを解析する（ステップ４０２）。ここでリードコマンドであることが分かるので（ステップ４０３）二次記憶装置２のローカルファイル管理テーブル２７の、後述するアクセス回数等の管理情報を必要に応じて更新し（ステップ４１１）、ディスク装置２９から求めるデータをリードし、上位計算機１にインタフェース制御手段２２を介し転送する（ステップ４１２）。そして、コマンド管理手段２３は、要求されたデータをすべて正しく転送できたならばインタフェース制御手段２２を介し上位計算機に終了報告を行う（ステップ４０８）。

【００６６】この終了報告は、上位装置１のデバイスドライバ１６によって受信される。

【００６７】図６に戻り、二次記憶装置２からのコマンド終了通知を受信すると（ステップ２０８）、デバイスドライバ１６は、これをファイル管理・バッファ管理手段１３に通知する。ファイル管理・バッファ管理手段１３は、転送すべきすべての論理ブロックのリードを完了したかどうかを確認し（ステップ２０９）、もしまだ未転送の論理ブロックが、あるならば続く論理ブロックを選択し（ステップ２１２）以上の処理を繰り返す。もしすべての論理ブロックのデータのシステムバッファへの転送が完了したならば、ＡＰが要求したオフセットデータをユーザバッファに複写し（ステップ２１０）、ＡＰにシステムコールの完了通知を返し、ファイルのリード処理を完了する。

【００６８】次に、ファイルのライト処理を、前述したリード処理との相違する点を中心に説明する。

【００６９】ファイルのライト処理は大きく次の２つの処理に分けられる。１つは新規ブロックの登録処理であり、もう１つは既存のブロックの更新処理である。前者は新規なファイルをライトする際や、既存ファイルの末尾へデータ追加を行う場合に発生し、後者は、既登録済み論理ブロックの途中からのデータをライトする場合や、既存ファイルをレコード（ひとまとまりのデータ単

17

位）単位で更新する場合等に発生する。

【0070】さて、図8に示すように、APからライトシステムコールを発行されると、（ステップ301）、OS20のファイル管理・バッファ管理手段13はディレクトリ管理テーブル14からファイル識別番号を獲得し（ステップ302）、前述したようにオフセットから転送すべき論理ブロックアドレスを求め（ステップ303）、必要なシステムバッファを確保する（ステップ304）。また、ここで、ファイル管理テーブル15を参照し、もし、論理ブロックアドレスに対応するローカルブロックアドレスが登録されていなかったならば、この論理ブロックは新規の論理ブロックであると判定する（ステップ305）。

【0071】新規の論理ブロックであると判定した場合には、ファイル管理・バッファ管理手段13は、ユーザバッファからライトデータをシステムバッファへ複写し（ステップ307）、最初の論理ブロックを選択する。一方、新規ファイルライト通知手段19は、選択した論理ブロックのライトを要求する新規ブロックライト要求を生成する。デバイスドライバ16は、これより、二次記憶装置2固有の新規ブロックライトコマンドを生成し、発行する（ステップ312）。

【0072】この新規ブロックライトコマンドにはファイル識別番号と論理ブロックアドレスが記載してある。

【0073】さて、発行された新規ブロックライトコマンドは、図9に示すように、二次記憶装置2のコマンド管理手段23で受領（ステップ401）され、解析される（ステップ402）。そして、解析の結果、新規ブロックライトコマンドであることを認識する（ステップ403、404）と、コマンド管理手段23はファイル固有識別番号および論理ブロックアドレスを新規ファイルアドレス決定手段261に渡す。新規ファイルアドレス決定手段26はファイル固有識別番号によりローカルファイル管理テーブル27およびディスク管理テーブル25bを参照し、このブロックを記憶するローカルブロックアドレスを、二次記憶装置2のディスク装置29の構成パラメータ、物理パラメータ等を考慮して決定する（ステップ405）。

【0074】このような最適なローカルブロックアドレスの決定手順については後述する。

【0075】次に、ローカルファイル管理テーブル27に、この決定したローカルブロックアドレスを論理ブロックアドレスに対応付けて登録し、また、ディスク管理テーブル25bに、当該ローカルブロックが使用中である旨設定する。

【0076】次に、データ転送を開始し、上位計算機1からデータの転送を受け、これを決定したローカルブロックアドレスに従いディスク装置29へライトする（ステップ406）。次に新規ファイルアドレス通知手段28は決定したローカルアドレスを上位計算機2に送信する（ステップ407）。そして、このブロックのローカルブロックアドレスの決定とデータ転送が終了したら終了通知を上位計算機1に行う（ステップ408）。

18

【0077】さて、送信されたローカルブロックアドレスは、上位装置において、インタフェース制御手段21及びデバイスドライバ16を介して、新規ファイルアドレス登録手段15に受信される。

【0078】図8に戻り、上位計算機1の新規ファイルアドレス登録手段15は、インタフェース制御手段21及びデバイスドライバ16を介して二次記憶装置2からローカルブロックアドレスを受信すると、ファイル管理・バッファ管理手段13にこのローカルブロックアドレスの登録を依頼する。ファイル管理・バッファ管理手段13は、ファイル管理テーブル15の、対応するファイルのファイル管理情報の対応する論理アドレスに対応付けて、このローカルブロックアドレスを登録する（ステップ314）。また、ファイル管理・バッファ管理手段13は、二次記憶装置2からのブロック転送終了通知を受信すると（ステップ315）、当該ブロックのライト処理を終了する。そして、すべての論理ブロックのライトが完了していないならば、次の該当論理ブロックを選択し、以上の処理をすべての論理ブロックのライト処理が終了するまで繰り返す（ステップ318）。

【0079】一方、ステップ305でライトするブロックが、新規論理ブロックでないと判定された場合、すなわち既存論理ブロックへのライト処理は、次のようになる。

【0080】この処理と、新規論理ブロックのライト処理との基本的な相違点は、既存論理ブロックへのライト処理では、ブロックの途中から更新するような場合がある点と、ライトする論理ブロックにはすでにローカルブロックアドレスがマッピングされている点である。

【0081】まず論理ブロックの途中から更新するような場合は、一旦、この論理ブロックをリードする必要がある。すなわち、システムバッファへ二次記憶装置から、ブロックをリードして、新データをユーザバッファからシステムバッファに転送して、システムバッファのブロックを更新し、その後、このブロックを二次記憶装置にライトするがある。

【0082】そこで、すなわち既存ブロックへのライトを行う場合には、まず、リード処理を行う（ステップ306）。このステップ306の処理は、図6中に示した符号aからbへの処理と等しい。

【0083】次に、既存のブロックへはすでにローカルブロックアドレスのマッピングが終了しているので二次記憶装置2へのライトコマンド発行時には、上位計算機1側でローカルブロックアドレスを指定する。このため、リード処理と同様の論理ブロックからローカルブロックアドレスへの変換処理を行い（ステップ310）、その後二次記憶装置への更新ライトコマンドを生成し発

行する（ステップ３１１）。

【００８４】一方、二次記憶装置２内は、受信した更新ライトコマンドに従い、指定されたローカルブロックアドレスに転送されたブロックのデータをライトする（図９ステップ４１０）。また、この際、二次記憶装置２のローカルファイル管理テーブル２７の、後述するアクセス回数等の管理情報を必要に応じて更新する（ステップ４０９）。

【００８５】以上の説明のように、本方式によれば、ファイルの二次記憶装置への格納に際しては、その格納するローカルアドレスを、二次記憶装置が決定するので、二次記憶装置特有の構成パラメータや、物理パラメータや、上位計算機１のＡＰの特性により決定されるファイルアクセス特性に合致した最適位置にファイルを格納を実現することができ、ファイルの高速化に効果が大きい。

【００８６】ところで、上位計算機１には、本第１実施例で説明してきたような二次記憶装置２の他、新規ファイルのローカルブロックアドレスを決定する能力がない従来の二次記憶装置２（図２４参照）をも接続できることが望ましい。すなわち、上位計算機１を、先に示した従来の二次記憶装置をも互換的に使用できるよう構成することが望ましい。

【００８７】そこで、本発明の第２実施例として、このような従来の二次記憶装置をも互換的に使用できる計算機システムについて説明する。図１０に、本第２実施例に係る計算機システムの構成を示す。

【００８８】本第２実施例に係る計算機システムは、本第１実施例に係る上位計算機１に、ディスク管理テーブル２５ａと、ファイルアドレス決定手段３０、ファイルアドレス決定手段３０と新規ファイルアドレス決定手段２６のどちらを使用するかを選択するアドレス決定手段選択手段３１を付加したものである。

【００８９】ディスク管理テーブル２５ａは、ディスクの使用状況を管理する。ファイルアドレス決定手段３０は、論理ブロックを新規にライトする際に書き込みに最適なディスク装置のローカルブロックアドレスを決定する。

【００９０】次に、本第２実施例に係る計算機システムの動作を説明する。

【００９１】まず、システムのイニシャル時等に、上位計算機１のファイル管理バッファ管理手段１３はデバイスドライバ１６を介して、二次記憶装置２のコマンド管理手段２３と、ネゴシエーションを行い、二次記憶装置２に新規ファイルのローカルブロックアドレスを決定する能力があるか否かを判定し、判定結果をアドレス決定手段選択手段３１に通知する。

【００９２】そして、アドレス決定手段選択手段３１は、もし、二次記憶装置２に新規ファイルのローカルブロックアドレスを決定する能力があれば、上位計算機１

のファイルアドレス決定手段３０とディスク管理テーブル２５ａを無効化とする。そして、以降は、前記第１実施例と同様の動作を行う。

【００９３】一方、アドレス決定手段選択手段３１は、もし、二次記憶装置２に新規ファイルのローカルブロックアドレスを決定する能力がなければ、オープンファイル情報通知手段１８、新規ファイルライト通知手段１９を無効化し、先に説明した従来の計算機システムと同様の動作を行う。

【００９４】さらに、複数二次記憶装置を同時に接続して使用するような場合は、ＡＰよりの二次記憶装置へのアクセスの要求毎に、要求された二次記憶装置が新規ファイルのローカルブロックアドレスを決定する能力があるか否かに応じて、オープンファイル情報通知手段１８と新規ファイルライト通知手段１９の組とファイルアドレス決定手段３０とディスク管理テーブル２５ａの組の一方の組を無効化し、前記第１実施例の動作もしくは従来の動作と同様の動作を行う。

【００９５】なお、二次記憶装置２に新規ファイルのローカルブロックアドレスを決定する能力がある場合でも、オープンファイル情報通知手段１８、新規ファイルライト通知手段１９を無効化すると共に、二次記憶装置２の新規ファイルアドレス決定手段２６、新規ファイルアドレス通知手段２８等を無効化することにより、従来と同様の動作を行うことができる。

【００９６】以上のように、本第２実施例によれば、前記第１実施例の計算機システムにおいて、従来の二次記憶装置をも利用することができる。また、複数台の二次記憶装置の接続や、従来使用していた二次記憶装置に新規の二次記憶装置を付設等を行うことができ、計算機システムの構成の自由度、拡張性を増すことができる。以下、本発明の第３の実施例を説明する。

【００９７】本第３実施例では、前記第１実施例と異なり、上位計算機１では、ファイルのローカルブロックアドレスを一切管理しない。

【００９８】図１１に、本第３実施例に係る計算機システムの構成を示す。

【００９９】図示するように、上位計算機１は、ディレクトリおよびファイル名とファイル識別番号を管理するディレクトリ管理テーブル１４と、データ転送に使用するバッファの管理を行うバッファ管理手段１３ａと、デバイスドライバ１６と、インタフェース制御手段２１とを備えている、また、二次記憶装置２は、インタフェース制御手段２２と、ファイル管理とコマンド管理を行うファイル管理・コマンド管理手段２３ａと、新規のファイルの論理ブロックへローカルブロックアドレスのマッピングを行う新規アドレス決定手段２６と、ファイルの管理情報を登録するローカルファイル管理テーブル２７と、ディスクの使用状況を管理し、またディスク装置２９の構成パラメータや、物理パラメータを記述しておく

21

ディスク管理テーブル２５ｂと、ディスク制御手段２４と、複数のディスク装置とを備えている。

【０１００】上位計算機１と二次記憶装置２とは、ファイルのオープンを上位計算機１から二次記憶装置２へ通知するコマンドや、ファイル識別番号とともにファイル中の指定のデータの転送を要求するコマンドをサポートしている。また、上位計算機１と二次記憶装置２とは、相互に、要求に見合った形態（ブロック単位転送またはバイト単位転送）でデータを転送することができる。このようなデータの転送や、コマンドの送受信は、インタフェース制御手段２１、２２が制御する。

【０１０１】以下、本第３実施例に係る計算機システムの動作を説明する。

【０１０２】プロセス１２は、ＯＳ２０に対しファイルのオープンや、ファイルのリード・ライトなどのシステムコールを発行する。ＯＳ２０はこれを受信し、バッファ管理手段１３ａはデータ転送に必要なバッファを確保し、デバイスドライバ経由で二次記憶装置２へのオープン、またはリード・ライトコマンドを生成し発行する。コマンドには、システムコールに含まれているオフセットを含める。また、このコマンドの発行に当たってはディレクトリ管理テーブル１４を参照し、このファイルのファイル識別番号を獲得し、コマンドとともに二次記憶装置２へ送信する。また、未登録ファイルのオープン時には、ファイル識別番号を新規に割り当て、割り当てたファイル識別番号を、コマンドとともに二次記憶装置２へ送信する。

【０１０３】二次記憶装置２において、ファイル管理・コマンド管理手段２３ａは、コマンドを受信、解析し、このコマンドが、新規の論理ブロックの登録を要求するコマンドであれば、新規ファイルアドレス決定手段２６に記憶アドレスの決定を依頼する。新規ファイルアドレス決定手段２６は、ディスク管理テーブル２５ｂ、ローカルファイル管理テーブル２７を参照し、ディスク装置２９の構成パラメータ、物理パラメータや、ＡＰのアクセス特性等を考慮し、この新規な論理ブロックを記憶するローカルブロックアドレスを決定し、ローカルファイル管理テーブル２７に登録する。ローカルブロックアドレスを決定の詳細については後述する。

【０１０４】また、ファイル管理・コマンド管理手段２３ａは、解析したコマンドが、既存の論理ブロックのリードもしくはライトを要求するコマンドであれば、ローカルファイル管理テーブル２７を参照し、アクセスする論理ブロックのローカルブロックアドレスを求め、このローカルブロックアドレスとコマンド中のオフセットで指定されるデータにアクセスする。そして、リードコマンドに対しては読みだしたデータを上位計算機に転送し、ライトコマンドに対しては転送されたデータをライトする。上位計算機１と二次記憶装置２との間のデータの転送は、ＡＰが要求した単位で行う。

22

【０１０５】以上のように、本第３実施例によれば、上位計算機は二次記憶装置の構成、物理的特性等をまったく考慮する必要がなく、ただファイル名とファイル識別番号のみを管理すればよい。したがい、様々な二次記憶装置や、その他の外部機器を上位計算機に接続して利用することができる。

【０１０６】以下、本発明の第４の実施例について説明する。

【０１０７】図１２に、本第４実施例に係る計算機システムの構成を示す。

【０１０８】本第４実施例に係る計算機システムは、前記第３実施例に係る上位計算機と二次記憶装置２を複数台接続したものである。

【０１０９】各上位計算機１と各二次記憶装置２は共通のインタフェースで接続されている。このインタフェースは、たとえばＳＣＳＩバスとして知られているバス型のインタフェースである。

【０１１０】本第４実施例に係る計算機システムの動作は、前記第３実施例とほぼ同様である。ただし、各ファイルの管理情報は各二次記憶装置毎に分散して配置されているため、各上位計算機１は、各ファイルを記憶している二次記憶装置を管理する。これは、各上位計算機が適当な契機で、各二次記憶装置から記憶しているファイルの情報を得ることにより実現できる。また、同じ二次記憶装置の異なるファイルに、異なる上位計算機によって同じファイル識別番号が割り当てられないように、二次記憶装置側で、新規ファイルへのファイル識別番号の二重割り当てを監視、調停する。これは、たとえば、上位装置よりの新規ファイルのオープン要求時に、ファイル識別番号の二重割り当てが発生した場合に、二次記憶装置が上位計算機にファイル識別番号の再割り当てを要求すること等により実現できる。さて、前記パスを介した、コマンド、データの転送のシーケンスは、たとえば、図１３に示すように実行される。すなわち、まず、アービトレーションを実行して、特定の上位計算機１と特定の二次記憶装置２を論理的に接続し、コマンドを上位計算機１より二次記憶装置に転送する。そして、次に、上位計算機は、二次記憶装置２のコマンドに対する前処理期間（ディスク装置のシーク期間等）中パスを他の装置に解放するため、メッセージ１を送信し一旦パスを解放する。前処理が終了すると、二次記憶装置２はメッセージ２を送信し、上位計算機とパスを再接続する。そして、次に、リード／ライトするデータを上位計算機１と二次記憶装置２間で転送し、最後に、メッセージ３によってコマンドの正常終了もしくは異常終了を二次記憶装置より上位計算機に通知し、パスを解放する。

【０１１１】なお、このような、パスは前記第１、第２実施例にも用いることができる。前記第１、第２実施例にも用いる場合には、メッセージ３によって新規な論理ブロックを記憶したローカルブロックアドレスも二次記

23

憶装置2から上位計算機1に転送するようにする。

【0112】以上のように、本第4実施例によれば、各上位計算機1はファイル名と、そのファイル識別番号、ファイルを記憶している二次記憶装置の識別以外の情報を保有する必要がないので、複数の上位計算機と複数の二次記憶装置を容易に接続できる。また、ファイル管理情報は、そのファイルを記憶している二次記憶装置のみが保有するので、二次記憶装置側で容易に、複数の上位装置に対するファイルのコヒレンシーを保つための制御を行うことができる。

【0113】さて、ここで、前述した二次記憶装置におけるファイルの各論理ブロックのローカルブロックアドレスの決定手順の詳細について説明する。

【0114】なお、以下に説明する手順は、前記第1実施例から第4実施例にいずれにも適用することができる。

【0115】さて、図14において、24はディスク装置制御手段、29は複数台で構成したディスク装置群である。ディスク装置制御手段24はディスク装置群29をRAID5型ディスクアレイとして管理している。RAID5型ディスクアレイは高速ディスクアクセスと高信頼性を目的としたディスクアレイアーキテクチャであり、このアーキテクチャでは、図示するようにデータをストライプと呼ぶ単位に分割し各ディスク装置に分散配置する。これによりストライプ単位のアクセスでは各ディスク装置を完全に独立に動作させることができ、ディスク台数倍のトランザクション性能を得られる。

【0116】さらに、横ならびのデータストライプ群（たとえば、D00、D01、D02）でパリティグループを構成し、これらデータの排他的論理和を計算しパリティ（たとえば、P0）として一台のディスク装置上に格納する。このパリティは各パリティグループ毎に異なるディスク装置に分散配置する。これにより、万一、1台のディスク装置が故障してもパリティを利用し故障ディスクのデータを復元できる。

【0117】このようなRAID5型ディスクアレイにおいて、各論理ブロックを最適に配置するためには、ストライプサイズや、ディスク台数や、パリティ配置方式等の構成パラメータが必要となる。しかし、これらのパラメータは、二次記憶装置の種別毎に異なるため、上位計算機において管理することは困難である。

【0118】一方、これらのパラメータを考慮しないと、例えばストライプサイズと論理ブロックサイズを同一サイズとしたとき、論理ブロックが1つのストライプの先頭、すなわちストライプバウンダリできちんと配置されず2つのストライプにまたがって配置されてしまう事態が発生する。そして、この場合、1つの論理ブロックをアクセスする際にも2つのストライプ、すなわち2台のディスク装置を同時にアクセスする必要が発生するので性能が著しく低下する。

24

【0119】そこで、本実施例では、二次記憶装置2側が自分の構成パラメータに基づいて論理ブロックの配置を決定する。以下、構成パラメータを考慮せずに行った論理ブロックの配置によって生じる問題と、これを回避するために二次記憶装置2が自分の構成パラメータに基づいて行う論理ブロックの配置の例を示す。

【0120】まず、第1の例を図15を用いて明する。

【0121】図15aは、ストライプ境界を意識せず論理ブロックの配置を決定した結果、データがディスク0のD_{00}ストライプとディスク1のD_{01}ストライプの2つのストライプを使用して記憶されることとなった場合を示している。この場合、1回の上位計算機からの要求に2台のディスク装置が使用されているので、これと並行して他の要求の処理に使用することができるディスクはディスク2，3の2台となる。

【0122】そこで、本実施例の二次記憶装置は、自分の構成パラメータに基づいて、上位計算機から転送された論理ブロックの先頭をストライプバウンダリに合わせるよう配置を決定することにより、図15bに示すようにディスク1のD_{01}ストライプのみにデータ格納する。この場合、上位計算機が8KB固定で新規論理ブロックのライト要求を発行する場合には、最大4つの要求を同時に処理でき、処理スループットを向上することができる。

【0123】次に、第2の例を図16を用いて説明する。

【0124】図16aは48KBの新規な論理ブロックを、ストライプバウンダリのみを考慮して配置したところを示している。ここで、RAID5型ディスクアレイではパリティを生成する必要がある。必要となるパリティはP_0，P_1，P_2であり、これらは、「*」を排他的論理和を表すものとして以下の生成式に従い生成できる。

【0125】$P_0New=(D_{01}Old*D_{02}Old)*(D_{01}New*D_{02}New)*P_0Old$

$P_1New=D_{10}New*D_{11}New*D_{31}New$

$P_2New=D_{20}Old*D_{20}New*P_2Old$

ここで、添字Oldのついたデータ（たとえば、$D_{01}Old$，$D_{02}Old$，P_0Old，…等）はディスクにすでに書き込まれているデータであり、添字Newのついたデータ（$D_{01}New$，$D_{02}New$，…等）は今まさに書き込もうとするデータである。したがい、P_0の生成には$D_{01}New$、$D_{02}New$の書き込みに先立ち、ディスク1，2，3より$D_{01}Old$、$D_{02}Old$、P_0Oldをリードする必要があり、P_2の生成には$D_{20}New$の書き込みに先立ちディスク0，1より$D_{20}Old$、P_2Oldをリードする必要がある。このライトに先立つリード（リードモディファイライト処理、以下、「RMW処理」と記す）は、多大な処理時間が必要となる。

25

【0126】そこで、本実施例の二次記憶装置は、自分の構成パラメータに基づいて、次のような手順で配置を決定する。

【0127】すなわち、新規ファイルアドレス決定手段26はローカルファイル管理テーブル27を参照し、概略アドレスを決定する。概略アドレスとは、ファイルへの論理ブロックの追加を行う場合においては、このファイルが現在格納されているアドレスであり、新規ファイルの論理ブロックの記憶ならば、現在ちょうど使用しているディスクの領域のアドレスである。新規な論理ブロックの概略アドレスをこのように選ぶのは、現在ちょうど使用しているディスク領域のデータと関係あると推測されるデータは、現在ちょうど使用しているディスク領域のデータの近くに配置した方が、確率的にディスクのヘッド移動に要する時間が少なくなるからである。

【0128】次に、概略アドレスが決定したならば、新規ファイルアドレス決定手段26はディスク管理テーブル25を参照し詳細アドレスを決定する。

【0129】すなわち、概略アドレスの近傍で丁度48KB分すなわち6ストライプを、ディスク0のストライプから開始し、順番に新規論理ブロックを配置可能な領域を、新規ファイルを格納するアドレスとして決定する。この例の場合、ディスク0のストライプ#nからディスク3のストライプ#n＋1のパリティストライプを除く計6つの連続したデータストライプが未使用であることをストライプ管理テーブルを参照し知ることができる。

【0130】ディスク管理テーブルは、図17に示すようにさらにストライプ管理テーブルとセクタ管理テーブルにより構成されている。

【0131】ここで、ストライプ管理テーブル1700は、ストライプ毎に、そのストライプがデータストライプがパリティストライプか、使用中か未使用か一部使用かを登録するエントリを有しており、これより、各ストライプがデータストライプがパリティストライプか、又、ストライプ全域を使用中か未使用か、一部使用かをディスク毎に知ることができる。また各ディスクのストライプ毎に、セクタの使用状況を管理するセクタ管理テーブル1710が設けられている。図17に示したセクタ管理テーブル1710は、ストライプサイズ8KBのときのもので、この場合、1ストライプは512Bのセクタ16コにより構成されている。このような構成により、各セクタが使用中か未使用かを判断できる。

【0132】たとえば、ディスク#0のストライプ#n＋2（図中丸印付加）は属性“010”であり一部使用中のデータストライプであることがわかる。そこで同ストライプのセクタ管理テーブルを参照すると、同ストライプ内のセクタ番号0～7は使用中であり、8～15は未使用であることがわかる。そして、これより、上位計算機より転送された新規論理ブロックを図16bのようにD00からD13までの連続した領域に格納することが決定される。

26

【0133】さらに、この場合、生成する必要のあるパリティストライプはP0，P1の2つであり、各々以下の生成式により生成される。

【0134】

$P_0New = D_{00}New * D_{01}New * D_{02}New$

$P_1New = D_{10}New * D_{11}New * D_{13}New$

この生成式より、ライトに先立つデータのリードを行わなくともパリティを求めることができることが判る。

【0135】さて、ここでRMW処理はW処理の平均1．7倍処理時間がかかると仮定すると、図16aのように記憶する場合に比べ、RMW処理が不必要になった分、2．2倍の高速に、新規論理ブロックを記憶が行われることになる。

【0136】以上のように、本実施例によれば二次記憶装置において、RAID型ディスクアレイのパリティ配置の方式や、ストライプサイズや、ディスク数等のパラメータに対し、データ転送長や関連する他のファイルとの位置関係を考慮した最適ファイル格納が実現できる。

【0137】なお、以上ではRAID5型ディスクアレイにおける最適論理ブロック配置の一例を示したが、いかなる構成のディスク装置（群）を用いたとしても、その構成パラメータや、物理特性、さらに論理ブロックサイズや、APのアクセス特性に合致した最適論理ブロック配置を実現できる。また、同様にして、ディスクアレイ以外の形態のディスク装置や、又は光ディスクやテープデバイスや半導体形記憶装置等についても最適配置を実現することができる。

【0138】以下、本発明の第5の実施例について説明する。

【0139】本第5実施例は、前記各実施例に係る計算機システムに用いることのできる二次記憶装置2に関するものである。

【0140】本第5実施例に係る二次記憶装置2と、前記第1実施例で示した二次記憶装置（図3参照）との主要な相違は、本第5実施例に係る二次記憶装置2が複数のディスク装置群を備えている点である。

【0141】図18に、本第5実施例に係る二次記憶装置の構成を示す。

【0142】図中、36a1、36a2、36a3はディスク装置群、24a1、24a2、24a3は各ディスク装置群を制御するディスク装置制御手段、34はどのディスク装置群を使用するかを選択するディスク装置群選択手段、35はディスク装置群を切り替えるディスク装置群切り換え手段、25b1、25b2、25b3は各ディスク群の構成パラメータや、物理特性や、使用状況を管理するディスク管理テーブル、1800はディスク群更新制御手段である。他部は、図3において同符号を付して示した部位と同じ部位であるので説明を省略

する。

【0143】以下、本第5実施例に係る二次記憶装置の動作の概要を説明する。

【0144】二次記憶装置2のコマンド管理手段23は、上位計算機1からアクセス要求を受信し、その要求が新規論理ブロックをライトする要求である場合には、新規ファイルアドレス決定手段26にこれを通知し、新規ファイルアドレス決定手段26は、ローカルファイル管理テーブル27、および、各ディスク管理テーブル25b1、25b2、25b3を参照し、最適なディスク装置群を選択しローカルブロックアドレスをマッピングする。そして、これをディスク装置群選択手段34に通知し、ディスク装置群選択手段34はこれを受けディスク装置群切り換え手段35を操作し、選ばれたディスク装置群を選択する。その後の処理は、第1実施例と同様である。

【0145】新規ファイルアドレス決定手段26が行う、ディスク装置群の選択は、転送長や、ファイル自体のアクセス頻度の特性や、ディスク装置群の特性等を考慮して行う。より詳細な例は後述する。

【0146】さて、本第5実施例に係るディスク装置群は、具体的には、たとえば、図19に示すように構成することができる。

【0147】図19に示した例では、ディスク装置群1はRAID1型ディスクアレイ装置（またはミラーディスク装置）、ディスク装置群2はRAID3型ディスクアレイ装置、ディスク装置群3はRAID5型ディスクアレイ装置としている。

【0148】この場合、アクセス頻度の高いファイルや高性能を要求するファイルはRAID1型のディスク装置群1に保管するのがふさわしく、画像データのような大容量シーケンシャルファイルはRAID3型のディスク装置群2に保管するのがふさわしく、データベースアクセスなどのトランザクション用ランダムファイルはRAID5型のディスク装置群3に格納するのがふさわしい。

【0149】そこで、新規ファイルアドレス決定手段26は、ディスク装置群の選択を次のように行う。

【0150】まず、新規ファイルアドレス決定手段26は、上位計算機よりのコマンドが既存のファイルの既存の論理ブロックを更新を要求するものである場合には、既存の論理ブロックを格納しているディスク装置群を選択する。

【0151】一方、上位計算機よりのコマンドが新規な論理ブロックの登録を要求するものである場合には、新規ファイルアドレス決定手段26は、まずどのディスク装置群に、この論理ブロックを登録するかを、図20に示す手順に従い決定する。

【0152】すなわち、まず、書き込む論理ブロックの属するファイルが新規か既存かの判断を行い（ステップ2001）、新規のファイルであれば、ファイルの特性が未知であるので、ひとまずRAID1型ディスクアレイ（ディスク群1）を選択する（ステップ2006）。

【0153】一方、書き込む論理ブロックの属するファイルが既存のファイルであれば、そのファイルのランダムアクセス性についての判断を行う（ステップ2002）。この判断には、ローカルファイル管理テーブル27を用いて行う。

【0154】すなわち、図21に示すように、ローカルファイル管理テーブルの各ファイル管理情報の、参照カウントフィールド156に、アクセス特性判定用の5つのサブフィールドを設ける。ここで、第1のサブフィールドは前回参照した上位計算機からの要求のアドレス（ファイル先頭からのバイト数）を格納するフィールド1561、第2のサブフィールドは前回参照したときの転送長（バイト数）を格納するフィールド1562、第3のサブフィールドはランダムアクセス回数の頻度を示すフィールド1563、第4のサブフィールドはシーケンシャルアクセス回数の頻度を示すフィールド1564、第5のサブフィールドはこのファイルの全参照回数を格納するフィールド1565である。また、ランダムアクセス回数フィールド1563、シーケンシャルアクセス回数フィールド1564の更新は、ファイルのアクセスの度に、次の処理により行われる。すなわち、式

今回参照アドレス≦前回参照アドレス＋前回転送長＋α

を判定し、式が成立すれば、このアクセスはシーケンシャルと判断し、シーケンシャルアクセス回数サブフィールド1564に1を加算する。もし不成立ならばランダムアクセスと判断し、ランダムアクセス回数サブフィールド1563に1を加算する。ただし、「今回参照アドレス」は上位計算機のライト要求のファイル先頭からの位置（バイト数）、αはシーケンシャル性の判断基準用定数である。なお、もし、完全に連続したアクセスのみをシーケンシャルとしたいときは、α＝0に設定すればよく、もし、ある程度幅をもたせてシーケンシャル性を判断したいときはαをある値に設定すればよい。このようにしてファイル参照のたびにランダム／シーケンシャル性を判定していく。

【0155】さて、ステップ2002に戻り、既存のファイルのランダムアクセス性についての判断、すなわち、ランダムアクセス性が大きいか、小さいかの判断は、次のように行う。

【0156】すなわち、A、Bを適当な設定値として、もし、対応するファイル管理情報の、参照回数フィールド1565の値がAより大きく、ランダムアクセス回数フィールド1563の値／参照回数フィールド1565の値がB以上であれば、ランダムアクセス性が大と判定し、参照回数フィールド1565の値がAより大きく、ランダムアクセス回数フィールド1563の値／参照回数フィールド1565の値がB未満であればランダムア

29

クセス性が小と判定する。また、参照回数フィールド１５６５の値がAより小さければ、判定不能と判断する。これは、参照回数がある一定値Aより小さい場合には、ファイルは作成されたばかりであって、まだシーケンシャル／ランダム性の判断できないからである。

【０１５７】次に、この判断の結果、ランダム性大ならばＲＡＩＤ５型ディスクアレイ（ディスク群3）を選択し（ステップ２００４）、判定不能ならばＲＡＩＤ１型ディスクアレイ（ディスク群１）を選択する（ステップ２００６）。ここで、判定不能の場合に、ＲＡＩＤ１型ディスクアレイ（ディスク群１）を選択するのは、作成したばかりの、ファイルはＲＡＩＤ１型ディスクアレイ（ディスク群１）に格納されているはずであるからである。

【０１５８】なお、ステップ２００２では、ランダムアクセス比率＝ランダムアクセス回数／参照回数が定数B（O＜B≦１）より大きいかどうかを判定したが、シーケンシャルアクセス比率をシーケンシャルアクセス回数フィールド１５６４の値／参照回数フィールド１５６５の値を基準にして判定を行うようにしてもよく、また、両者を加味した判定を行ってもよい。

【０１５９】さて、ステップ２００２で、ランダム性が小と判断されたならば転送長の判定を行う。すなわち、コマンドで書き込みを要求されているデータ長（転送長）がある一定値C以上ならば、転送長大と判断し、ＲＡＩＤ３型ディスクアレイ（ディスク群２）を選択する。また転送長ががある一定値C未満のときは、ＲＡＩＤ３型は効率が悪いのでＲＡＩＤ１型を選択する。

【０１６０】以上のように、新規ファイルが既存ファイルか、ランダム性が大か小か、転送長が大か小かの判定を行うことにより、適切なディスク群を選択する。そして、選択したディスク群に、新規な論理ブロックを書き込むと共に、選択したディスク群が、新規な論理ブロックの属する既存のファイルの既存の論理ブロックが記憶されているディスク群と異なる場合には、選択したディスク群に既存の論理ブロックを移動し、これに合わせファイル管理テーブルを更新する。なお、この際、ローカルファイル管理テーブル２７の各ファイル管理情報に設けたファイルの再配置の有無を示す更新フラグ１８８を”更新済み”に設定する。

【０１６１】なお、以上では、新規ファイルはすべて、ＲＡＩＤ１型ディスクアレイ（ディスク群１）に書き込むこととした。これはアクセス特性が不明であるためである。しかしながらＲＡＩＤ１型ディスクアレイは信頼性が高い点を除いては１台のディスクと同等の性能であり、ランダム性の強いファイルや、シーケンシール性の強いファイルのアクセスには適さない。また、ディスク容量も限られているので、適当なタイミングでそのアクセス特性に適したディスク群を選択し直し、ファイルを移動させることが望ましい。

30

【０１６２】そこで、ディスク群更新制御手段１８００が、定期的に、コマンド管理手段２３にファイル毎に格納するディスク群を選択しなおすよう要求を発行するようにする。

【０１６３】コマンド管理手段２３はこの要求をうけ、ローカルファイル管理テーブル２７の各ファイル管理情報の更新フラグ１８８を参照し、もし、更新フラグが“未更新”に設定されていたならば最近新規に作成されたファイルなので、ディスク装置群選択手段を起動し、図２０に示した手順に従いディスク群の再選択を試みる。

【０１６４】そして、ランダム性又はシーケンシャル性が強いファイルであることが判定されたなら上述のようなディスク群を選択して、ＲＡＩＤ１型ディスクアレイ（ディスク群１）以外のディスク群が選択されたら、そのディスク群にファイルを移動し、更新フラグ１８８を“更新済”に設定する。

【０１６５】もし、判定不能であれば、そのままディスク群１に当該ファイルを置く。このとき、更新フラグは“未更新”のままにする。なお、ファイル移動中にこのファイルへのアクセス要求が上位計算機から発行されることがあるため、このとき更新フラグは“更新中”に設定しておく。また、ファイルの移動に伴いファイル管理テーブルを新規に作成し直し、ファイルの移動が完全に終了したら、ファイル管理テーブルを新規のものに切りかえ古いものは削除してしまう。なお、ディスク群の選択後の論理ブロックの論理ブロックアドレスへの登録は先に示したように行えばよい。

【０１６６】以上の処理により、アクセス特性に適したディスク群を選択できる。

【０１６７】なお、以上ではディスク群としてＲＡＩＤ１型、ＲＡＩＤ３型、ＲＡＩＤ５型のディスクアレイを使用した例を使用したが、これ以外の装置を導入しても、同様な手順によりアクセス特性に適したディスク群を選択できる。

【０１６８】さて、本第５実施例に係るディスク装置群（図１９参照）は、図２２に示すように構成することもできる。

【０１６９】図２２に示した例では、構成パラメータの１つであるストライプサイズがそれぞれ異なる３つのディスクアレイ装置をディスク装置群１、２、３としている。この場合、比較的小さなファイルを格納するにはストライプサイズを小さく設定したディスク装置群１に、また逆に十分大きなファイルを格納するにはストライプサイズを大きく設定したディスク装置群３に登録するのがふさわしい。したがい、ファイル長に応じて、ディスク装置群を選択すれば、ファイルの特性に適したディスク群を選択できる。

【０１７０】または、データの転送長に応じて、ストライプサイズ×ディスク台数×ｎが転送長となるようなデ

31

ィスク群を選択するようにすれば、効率よく、ディスク台数分の転送性能を得ることができる。

【０１７１】また、本第５実施例に係るディスク装置群（図１９参照）は、図２３に示すように構成することもできる。

【０１７２】図２３に示した例では、ディスク装置群１に磁気ディスク装置を、ディスク装置群２に光ディスク装置を、ディスク装置群３にテープライブラリ装置を用いている。この場合、アクセス頻度の高いファイルは磁気ディスク装置に、アクセス頻度の余り高くないファイルは光ディスク装置に、ほとんどアクセスしないファイル（バックアップ等）はテープライブラリ装置に登録することが望ましい。

【０１７３】そこで、ある一定時間毎に、ディスク群更新制御手段１８００が、コマンド管理手段２３にファイル毎に格納するディスク群を選択しなおすよう要求を発行するようにする。そして、コマンド管理手段２３が、この要求をうけ、ローカルファイル管理テーブル２７の各ファイル管理情報のタイムスタンプ１５３の前回参照時間と、参照回数フィールド１５６５を参照し、このファイルが最近アクセスされているのかどうか、又は参照回数が多いかどうかを判定する。そして、もし一定期間アクセスされていないならば古いファイルであり、高速なディスク上のスペースを占拠していることは非効率であるので、たとえば光ディスク等にファイルを移動するようにする。

【０１７４】なお、たとえば夜間に、テープライブラリ装置に、他のディスク装置に記憶されているファイルを複写し、バックアップを取るようにしてもよい。

【０１７５】以上のように本実施例によれば複数のディスク装置群を二次記憶装置内に設けることができ、その構成を変えることで多種多様なアクセス要求やファイル特性に適した論理ブロックの論理ブロックアドレスへのマッピングを実現でき、様々な形態の記憶システムおよび計算機システムを構築できるという効果がある。

【０１７６】以上説明してきたように、本実施例によれば、二次記憶装置固有の構成パラメータや、物理パラメータを配慮し、かつ上位計算機のＡＰのアクセス特性に合致した最適ファイル配置を実現でき、高速なファイルアクセスを実現できる。またこれにより高性能な計算機システムを実現できる。また、上位計算機に本発明の二次記憶装置のみならず従来の二次記憶装置をも接続でき、柔軟な計算機システムを構築できる。また、上位計算機は二次記憶装置の構成、物理特性を考慮する必要がないので、簡単に様々な二次記憶装置を上位装置に接続できる。また、複数台の上位計算機と複数台の二次記憶装置を接続した分散ファイル管理型計算機システムを容易に構築できる。

【０１７７】

【発明の効果】以上のように、本発明によれば、上位計算機に二次記憶装置のパラメータを設定すること無しに、ファイルの二次記憶装置への最適配置を実現することのできる計算機システムを提供することができる。

32

【図面の簡単な説明】

【図１】本発明の実施例に係る計算機システムのハードウェア構成を示すブロック図である。

【図２】本発明の実施例におけるファイル管理の概念を示した説明図である。

【図３】本発明の第１実施例に係る計算機システムの構成を示したブロック図である。

【図４】本発明の第１実施例に係るファイル管理テーブルの構成を示す説明図である。

【図５】本発明の第１実施例に係るＯＳが行うファイルオープン処理の手順を示すフローチャートである。

【図６】本発明の第１実施例に係るＯＳが行うファイルリード処理の手順を示すフローチャートである。

【図７】アクセスするデータと論理ブロックの関係を示した説明図である。

【図８】本発明の第１実施例に係るＯＳが行うファイルライト処理の手順を示すフローチャートである。

【図９】本発明の第１実施例に係る二次記憶装置が行うファイルリード／ライト処理の手順を示すフローチャートである。

【図１０】本発明の第２実施例に係る計算機システムの構成を示したブロック図である。

【図１１】本発明の第３実施例に係る計算機システムの構成を示したブロック図である。

【図１２】本発明の第４実施例に係る計算機システムの構成を示したブロック図である。

【図１３】本発明の第４実施例に係る上位計算機と二次記憶装置間の転送シーケンスを示すタイムチャートである。

【図１４】ディスクアレイ装置のストライプを示す説明図である。

【図１５】論理ブロックへのローカルアドレスの割り当てのようすを示す説明図である。

【図１６】論理ブロックへのローカルアドレスの割り当てのようすを示す説明図である。

【図１７】本発明の実施例に係るストライプ管理テーブルとセクタ管理テーブルの構成を示すブロック図である。

【図１８】本発明の第５実施例に係る二次記憶装置の構成を示すブロック図である。

【図１９】本発明の第５実施例に係る二次記憶装置の第１の具体的構成例を示すブロック図である。

【図２０】本発明の第５実施例に係るディスク群選択処理の手順を示すフローチャートである。

【図２１】本発明の第５実施例に係るローカルファイル管理テーブルの構成を示す説明図である。

【図２２】本発明の第５実施例に係る二次記憶装置の第

33

２の具体的構成例を示すブロック図である。

【図２３】本発明の第５実施例に係る二次記憶装置の第３の具体的構成例を示すブロック図である。

【図２４】従来の計算機システムの構成を示すブロック図である。

【符号の説明】

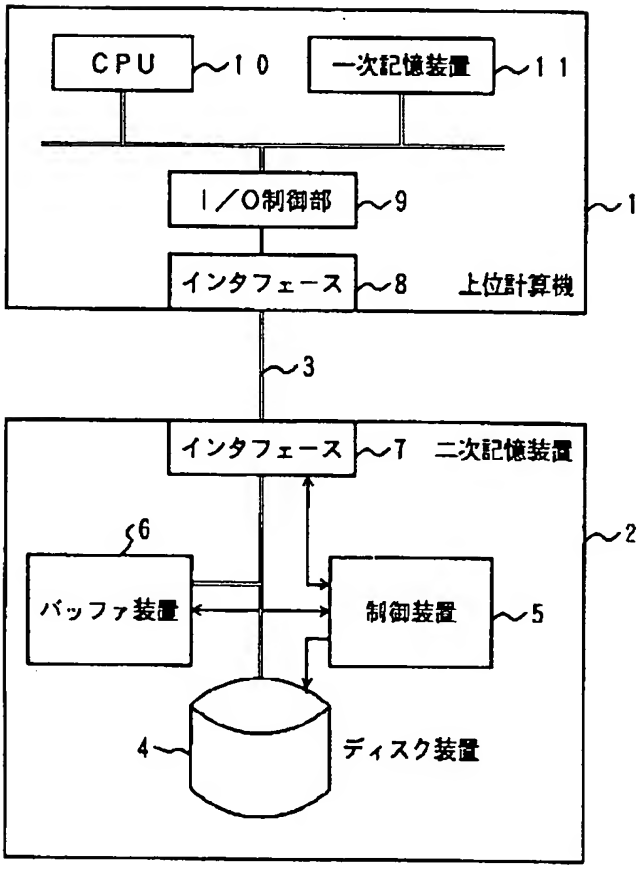
１・・・上位計算機
２・・・二次記憶装置
１２・・・プロセス
１３・・・ファイル管理・バッファ管理手段
１４・・・ディレクトリ管理テーブル
１５・・・ファイル管理テーブル
１６・・・デバイスドライバ

34

１７・・・新規ファイルアドレス決定手段
１８・・・オープンファイル情報通知手段
１９・・・新規ファイルライト通知手段
２０・・・オペレーティングシステム
２１・・・インタフェース制御手段
２２・・・インタフェース制御手段
２３・・・コマンド管理手段
２４・・・ディスク装置制御手段
２５ｂ・・・ディスク管理テーブル
２６・・・新規ファイルアドレス決定手段
２７・・・ローカルファイル管理テーブル
２８・・・新規ファイルアドレス通知手段
２９・・・ディスク装置

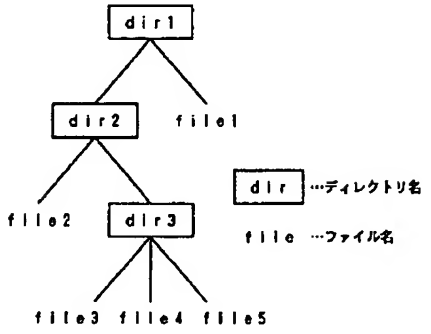
【図１】

図１



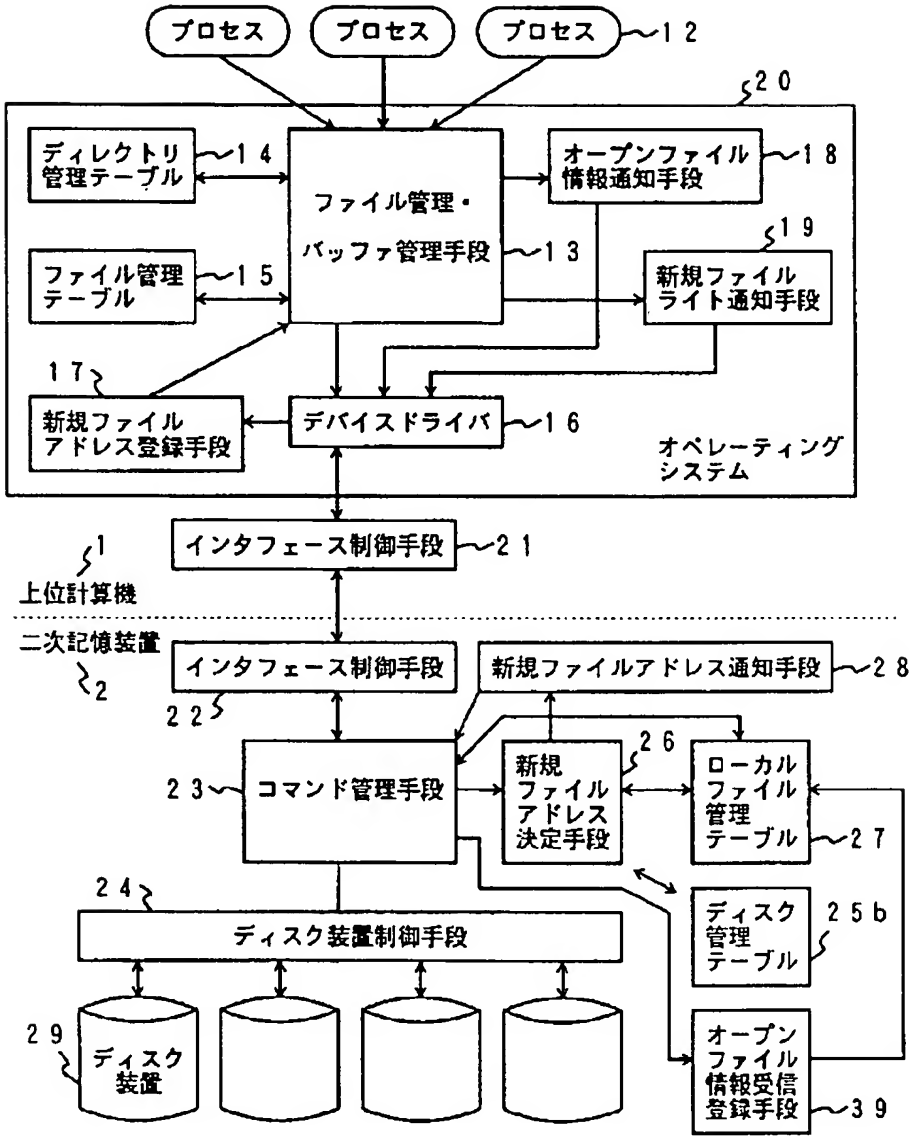
【図２】

図2



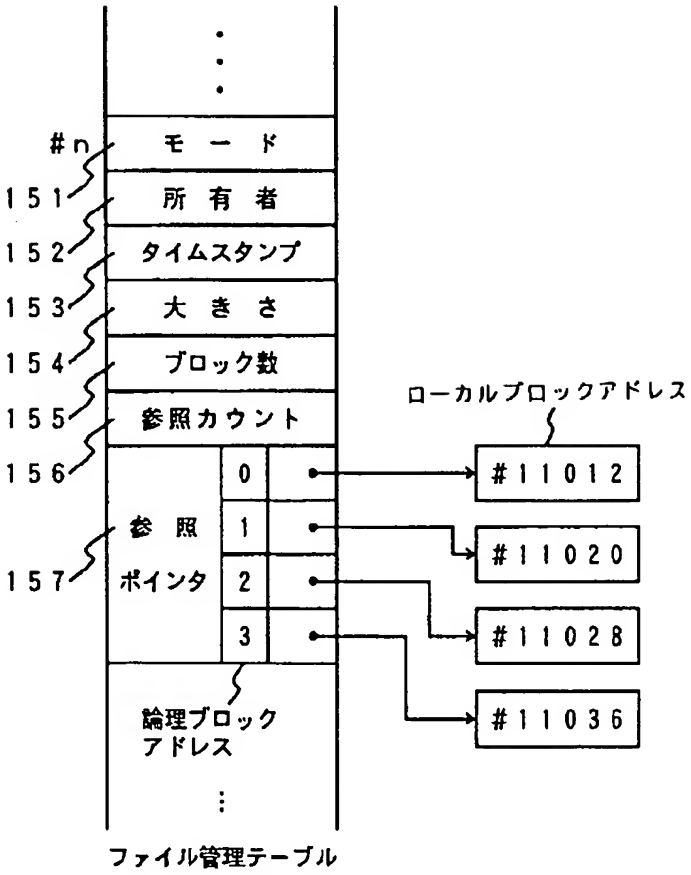
【図３】

図３


プロセス
プロセス
プロセス
12
20
ディレクトリ管理テーブル
14
ファイル管理・バッファ管理手段
13
オープンファイル情報通知手段
18
19
ファイル管理テーブル
15
新規ファイルライト通知手段
17
新規ファイルアドレス登録手段
デバイスドライバ
16
オペレーティングシステム
1
上位計算機
インタフェース制御手段
21
二次記憶装置
2
インタフェース制御手段
新規ファイルアドレス通知手段
28
22
26
23
コマンド管理手段
新規ファイルアドレス決定手段
ローカルファイル管理テーブル
27
24
ディスク装置制御手段
ディスク管理テーブル
25b
29
ディスク装置
オープンファイル情報受信登録手段
39

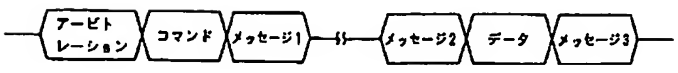
【図4】

図4


#n
モード
151
所有者
152
タイムスタンプ
153
大きさ
154
ブロック数
155
参照カウント
156
0
1
2
3
参照
ポインタ
157
論理ブロック
アドレス
ローカルブロックアドレス
#11012
#11020
#11028
#11036
ファイル管理テーブル

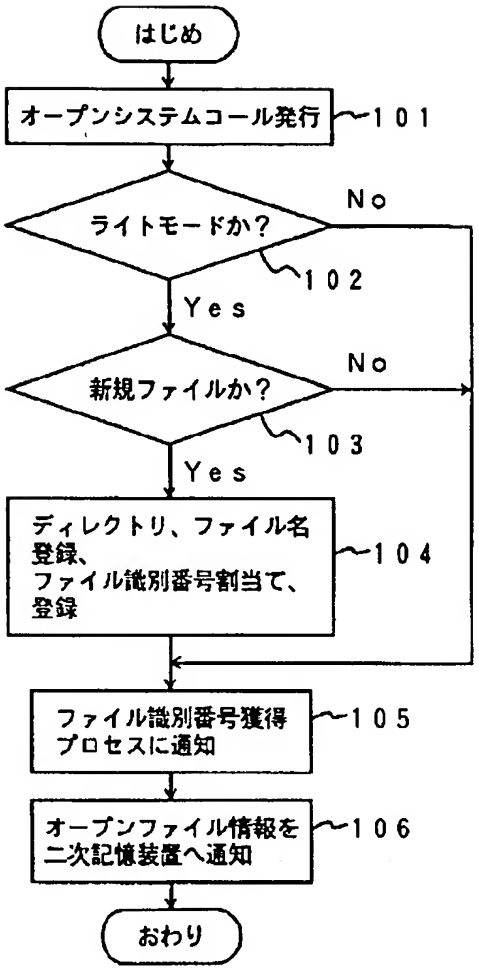
【図13】

図13


アービトレーション
コマンド
メッセージ1
メッセージ2
データ
メッセージ3

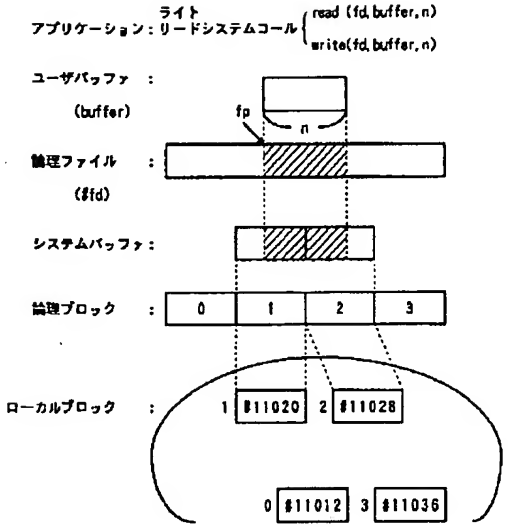
【図5】

図5



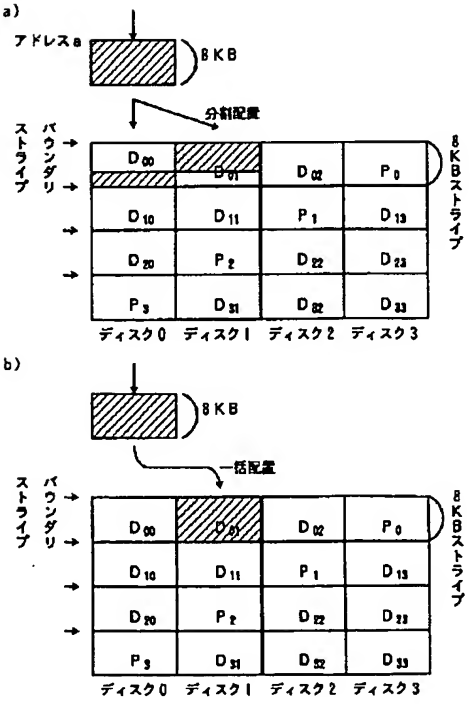
【図7】

図7



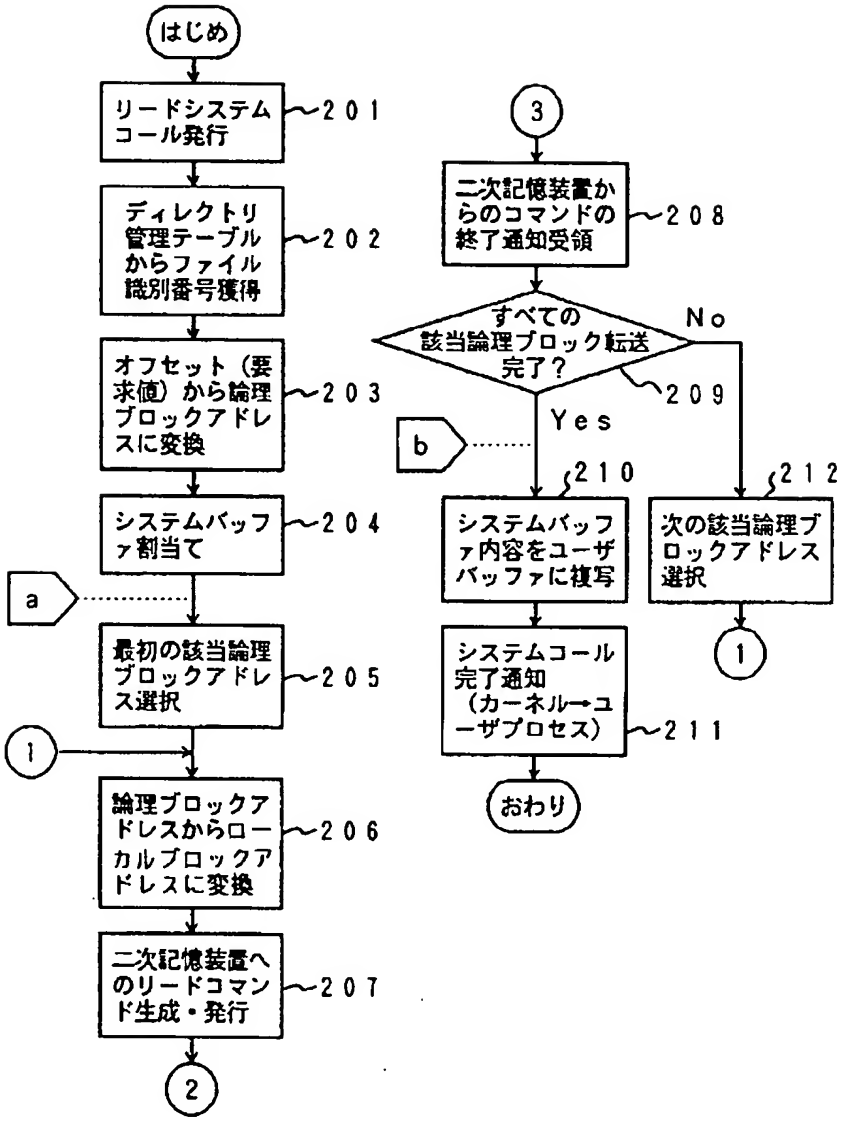
【図15】

図15



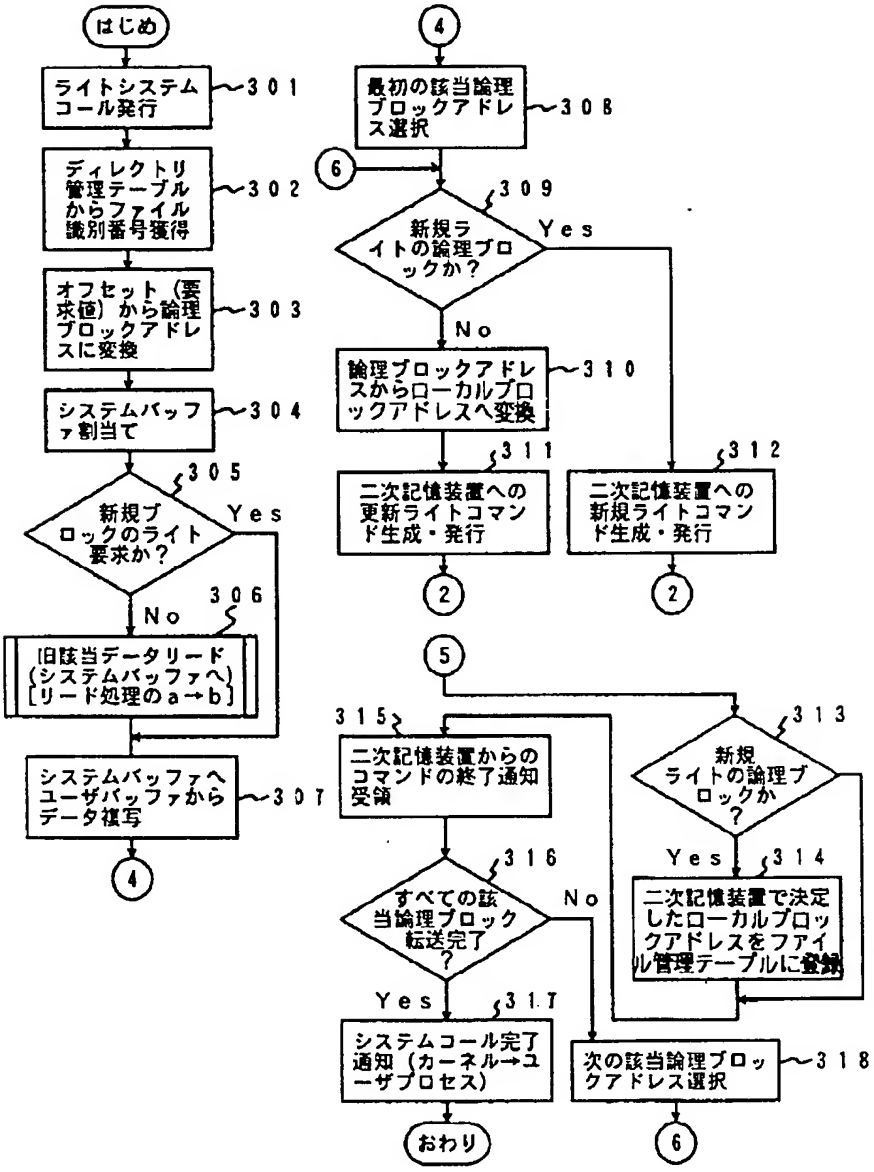
【図6】

図6


はじめ
リードシステムコール発行
201
ディレクトリ管理テーブルからファイル識別番号獲得
202
オフセット（要求値）から論理ブロックアドレスに変換
203
システムバッファ割当て
204
a
最初の該当論理ブロックアドレス選択
205
1
論理ブロックアドレスからローカルブロックアドレスに変換
206
二次記憶装置へのリードコマンド生成・発行
207
2
3
二次記憶装置からのコマンドの終了通知受領
208
すべての該当論理ブロック転送完了？
209
No
Yes
b
210
システムバッファ内容をユーザバッファに複写
システムコール完了通知（カーネル→ユーザプロセス）
211
おわり
212
次の該当論理ブロックアドレス選択
1

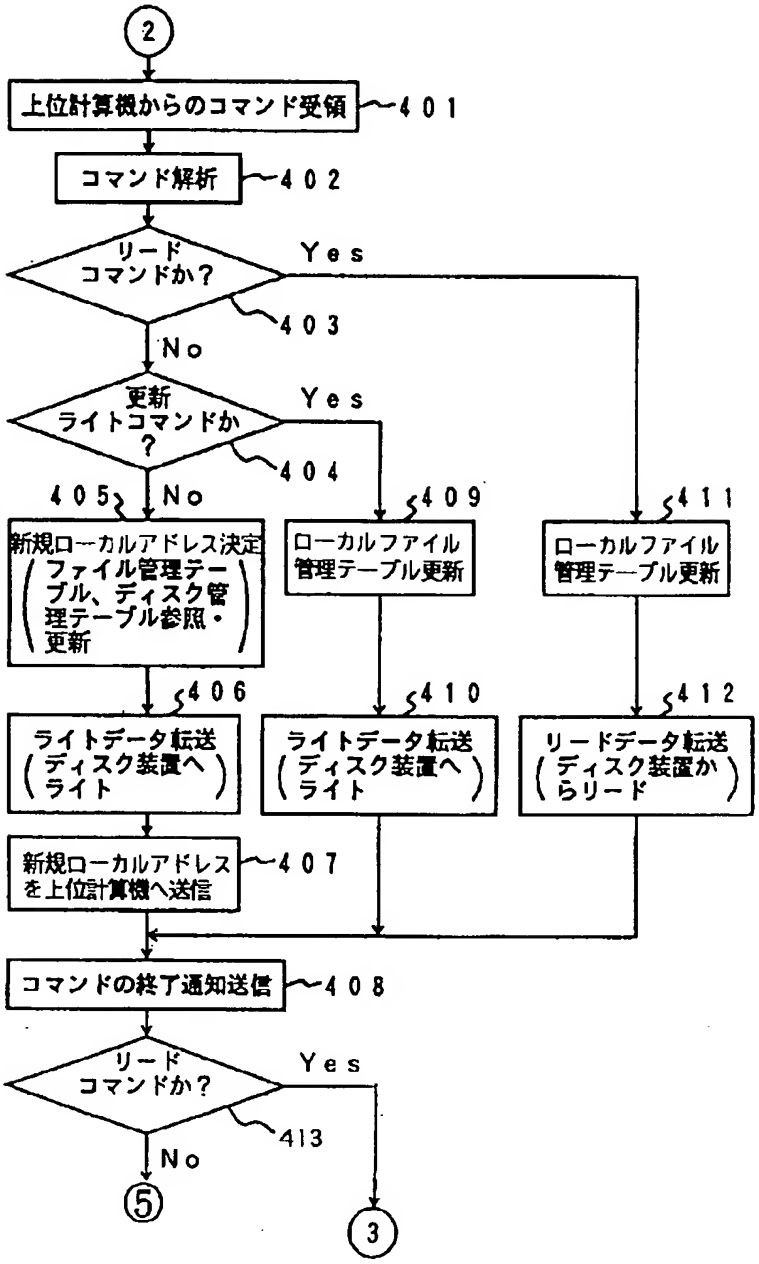
【図8】

図8


はじめ
ライトシステムコール発行
301
ディレクトリ管理テーブルからファイル識別番号獲得
302
オフセット（要求値）から論理ブロックアドレスに変換
303
システムバッファ割当て
304
新規ブロックのライト要求か？
305
Yes
No
旧該当データリード（システムバッファへ）［リード処理のa→b］
306
システムバッファへユーザバッファからデータ複写
307
4
4
最初の該当論理ブロックアドレス選択
308
6
新規ライトの論理ブロックか？
309
Yes
No
論理ブロックアドレスからローカルブロックアドレスへ変換
310
二次記憶装置への更新ライトコマンド生成・発行
311
二次記憶装置への新規ライトコマンド生成・発行
312
2
2
5
二次記憶装置からのコマンドの終了通知受領
315
新規ライトの論理ブロックか？
313
Yes
二次記憶装置で決定したローカルブロックアドレスをファイル管理テーブルに登録
314
すべての該当論理ブロック転送完了？
316
No
Yes
システムコール完了通知（カーネル→ユーザプロセス）
317
次の該当論理ブロックアドレス選択
318
おわり
6

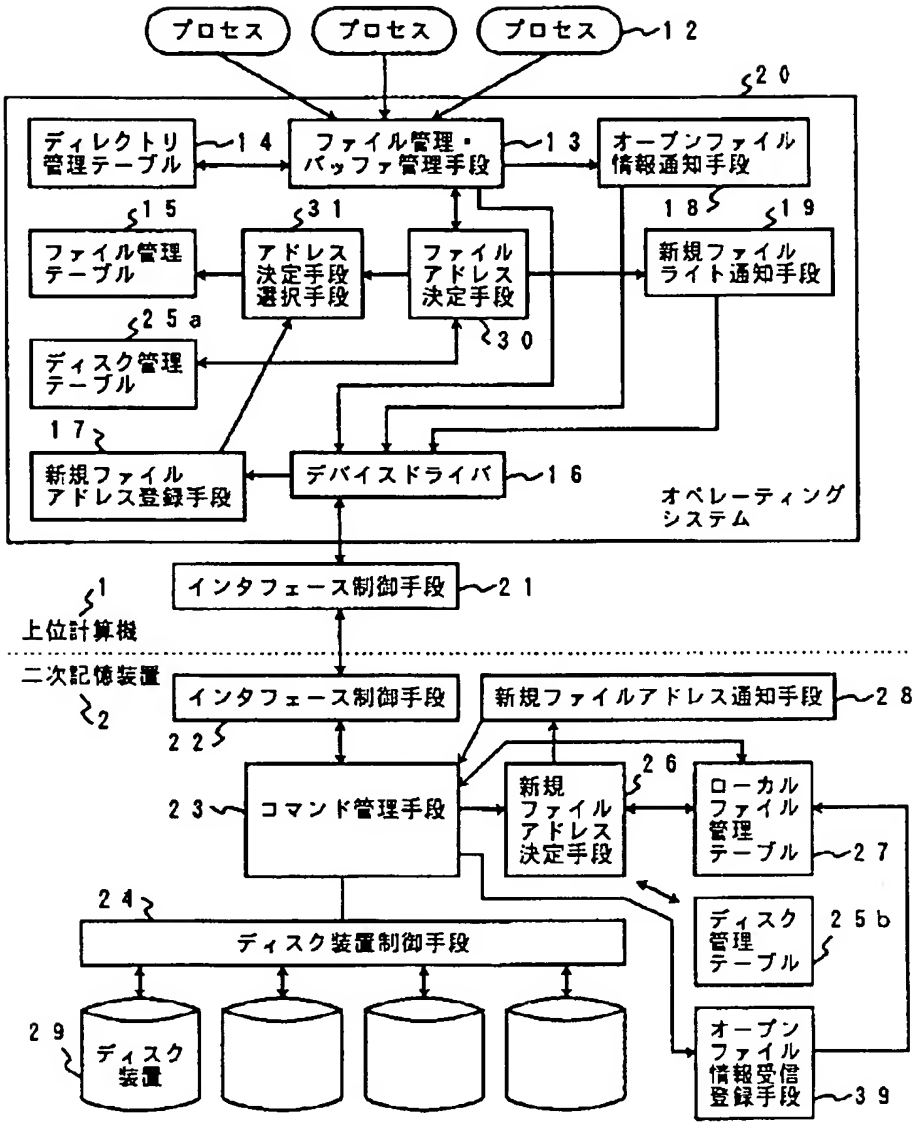
【図９】

図９


2
上位計算機からのコマンド受領
401
コマンド解析
402
リードコマンドか？
Yes
403
No
更新ライトコマンドか？
Yes
404
405
No
409
411
新規ローカルアドレス決定（ファイル管理テーブル、ディスク管理テーブル参照・更新）
ローカルファイル管理テーブル更新
ローカルファイル管理テーブル更新
406
410
412
ライトデータ転送（ディスク装置へライト）
ライトデータ転送（ディスク装置へライト）
リードデータ転送（ディスク装置からリード）
新規ローカルアドレスを上位計算機へ送信
407
コマンドの終了通知送信
408
リードコマンドか？
Yes
413
No
⑤
3

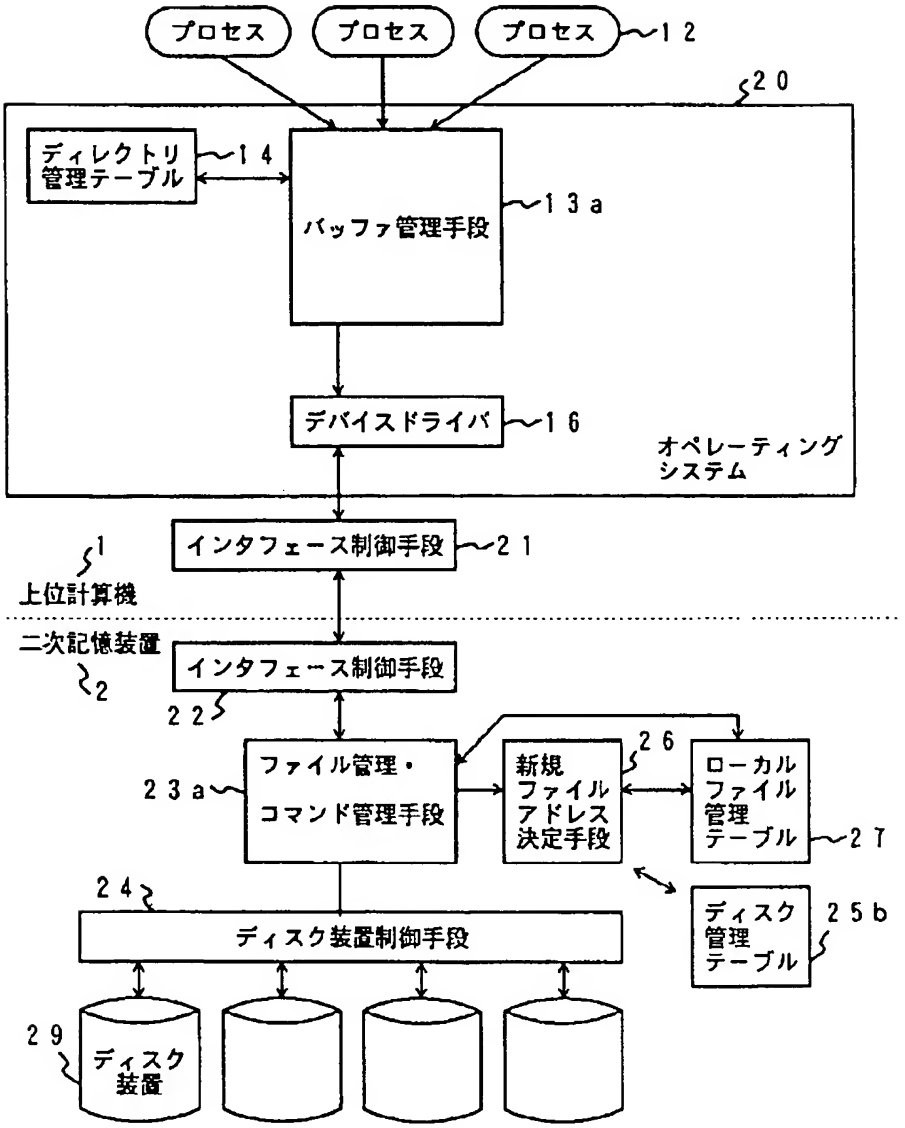
【図１０】

図10


プロセス
プロセス
プロセス
12
20
ディレクトリ管理テーブル
14
ファイル管理・バッファ管理手段
13
オープンファイル情報通知手段
15
ファイル管理テーブル
31
アドレス決定手段選択手段
ファイルアドレス決定手段
30
18
19
新規ファイルライト通知手段
25a
ディスク管理テーブル
17
新規ファイルアドレス登録手段
デバイスドライバ
16
オペレーティングシステム
1
上位計算機
インタフェース制御手段
21
二次記憶装置
2
インタフェース制御手段
22
新規ファイルアドレス通知手段
28
23
コマンド管理手段
新規ファイルアドレス決定手段
26
ローカルファイル管理テーブル
27
24
ディスク装置制御手段
ディスク管理テーブル
25b
29
ディスク装置
オープンファイル情報受信登録手段
39

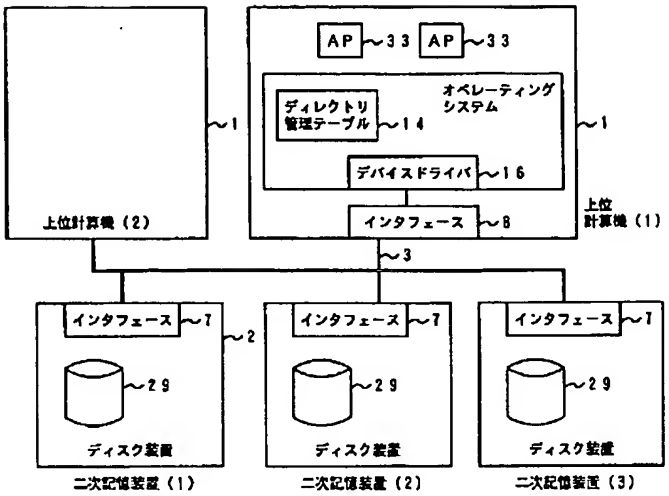
【図１１】

図１１


プロセス
プロセス
プロセス
１２
２０
ディレクトリ管理テーブル
１４
バッファ管理手段
１３ａ
デバイスドライバ
１６
オペレーティングシステム
１
上位計算機
インタフェース制御手段
２１
二次記憶装置
２
インタフェース制御手段
２２
２３ａ
ファイル管理・コマンド管理手段
２６
新規ファイルアドレス決定手段
ローカルファイル管理テーブル
２７
２４
ディスク装置制御手段
ディスク管理テーブル
２５ｂ
２９
ディスク装置

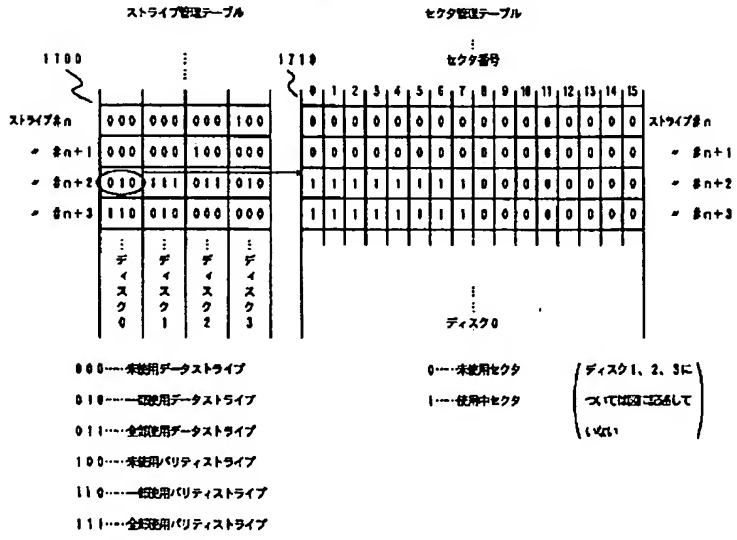
【図１２】

図12



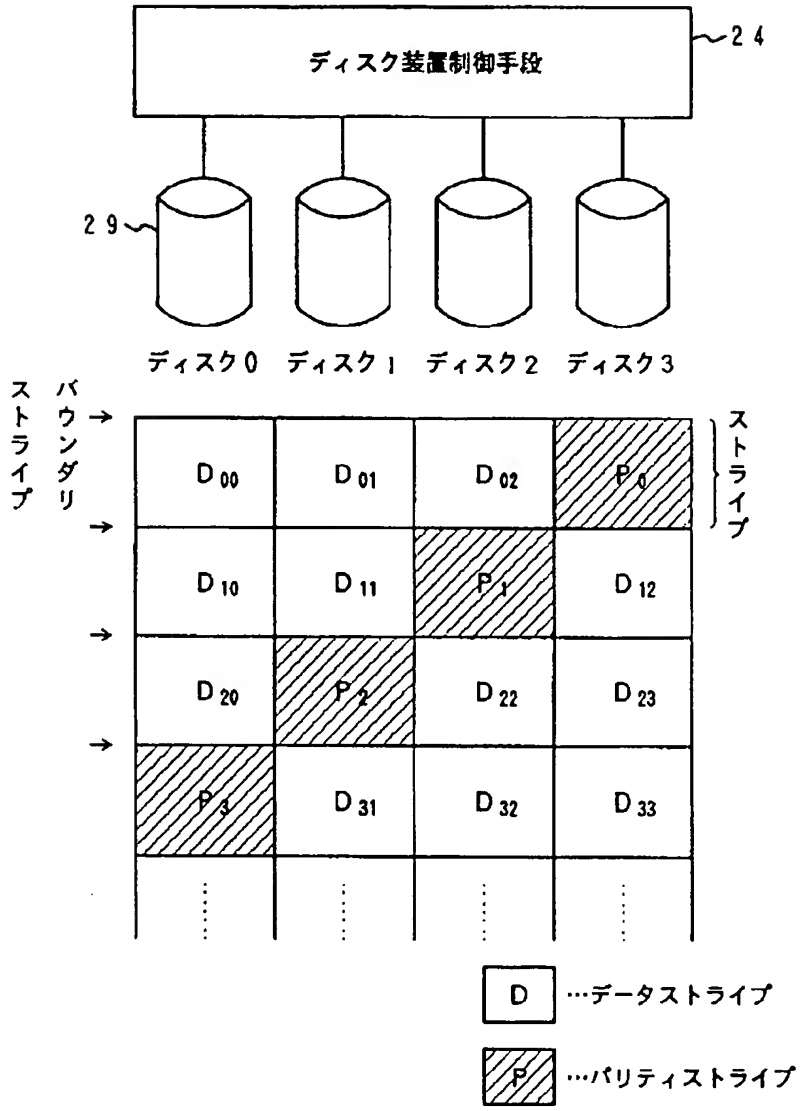
【図１７】

図17



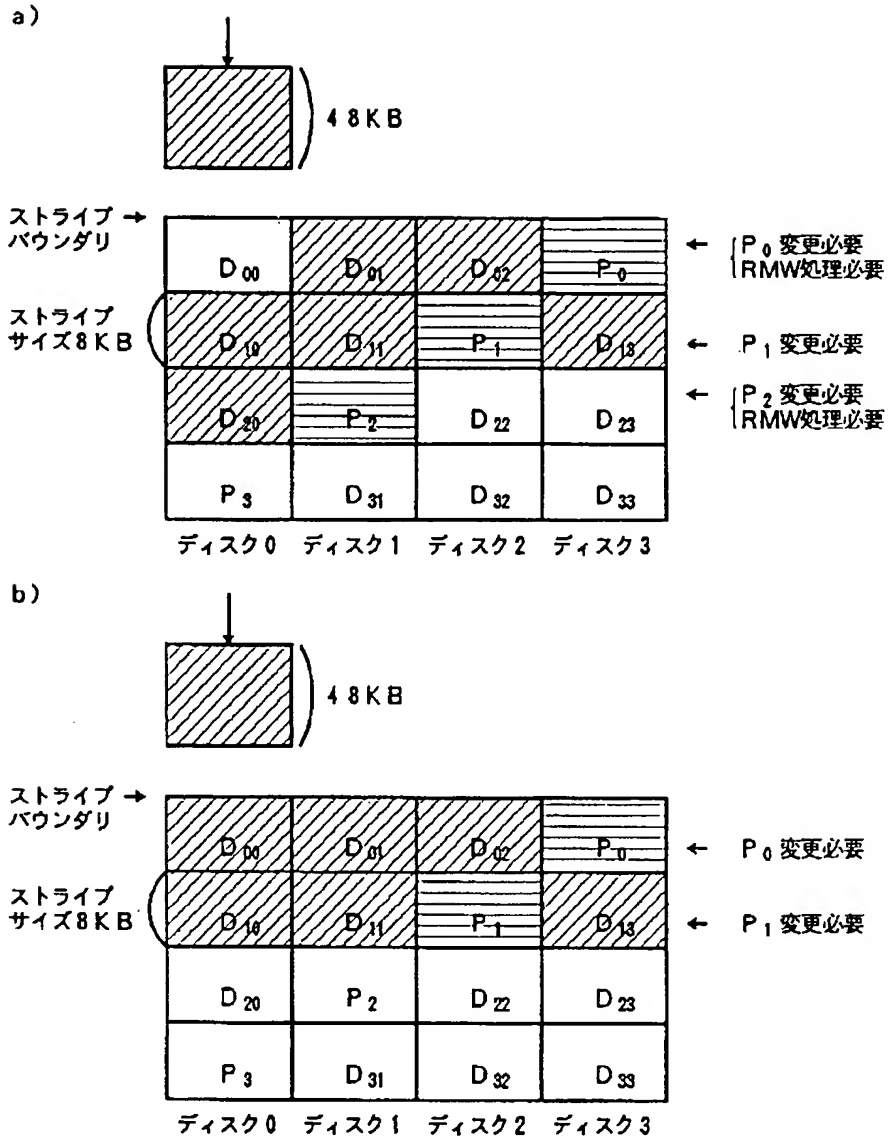
【図14】

図14


ディスク装置制御手段
24
29
ディスク0
ディスク1
ディスク2
ディスク3
バウンダリストライプ
ストライプ
D00
D01
D02
P0
D10
D11
P1
D12
D20
P2
D22
D23
P3
D31
D32
D33
D …データストライプ
P …パリティストライプ

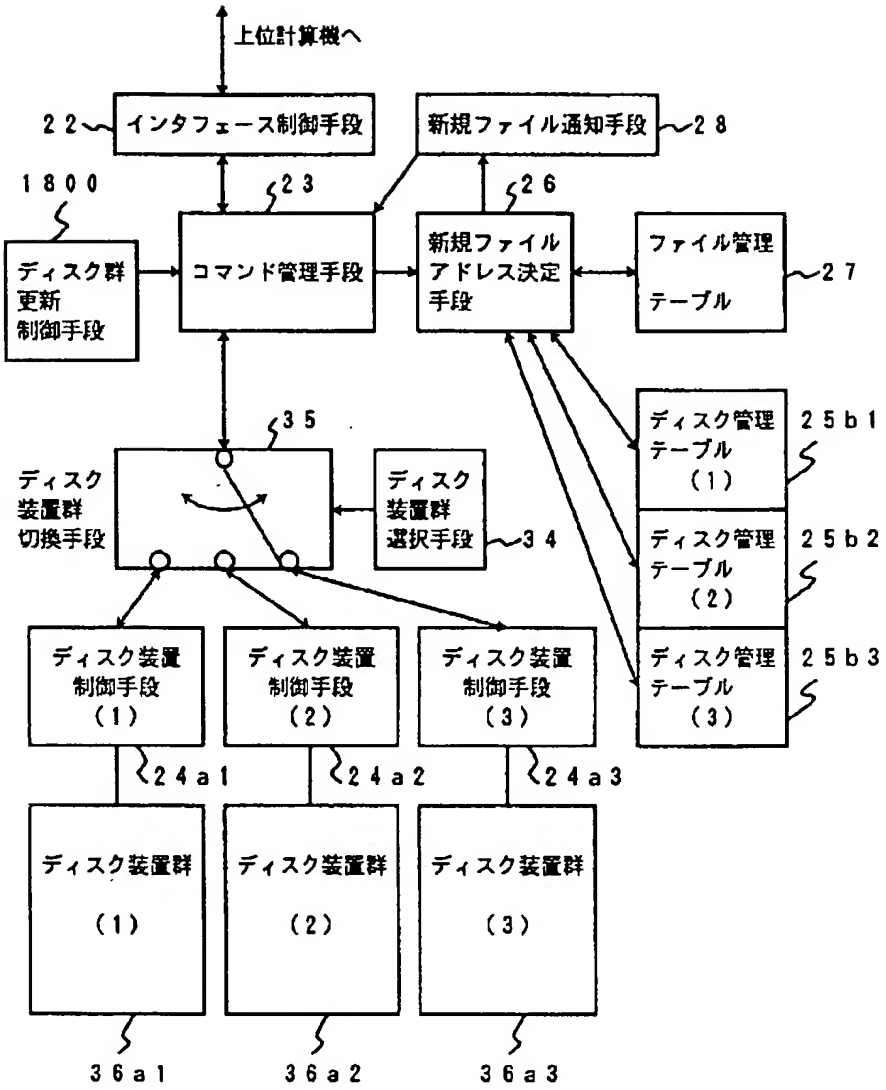
【図１６】

図16


a)
48KB
ストライプ
バウンダリ
ストライプ
サイズ8KB
D00
D01
D02
P0
D10
D11
P1
D13
D20
P2
D22
D23
P3
D31
D32
D33
P0 変更必要
RMW処理必要
P1 変更必要
P2 変更必要
RMW処理必要
ディスク0
ディスク1
ディスク2
ディスク3
b)
48KB
ストライプ
バウンダリ
ストライプ
サイズ8KB
D00
D01
D02
P0
D10
D11
P1
D13
D20
P2
D22
D23
P3
D31
D32
D33
P0 変更必要
P1 変更必要
ディスク0
ディスク1
ディスク2
ディスク3

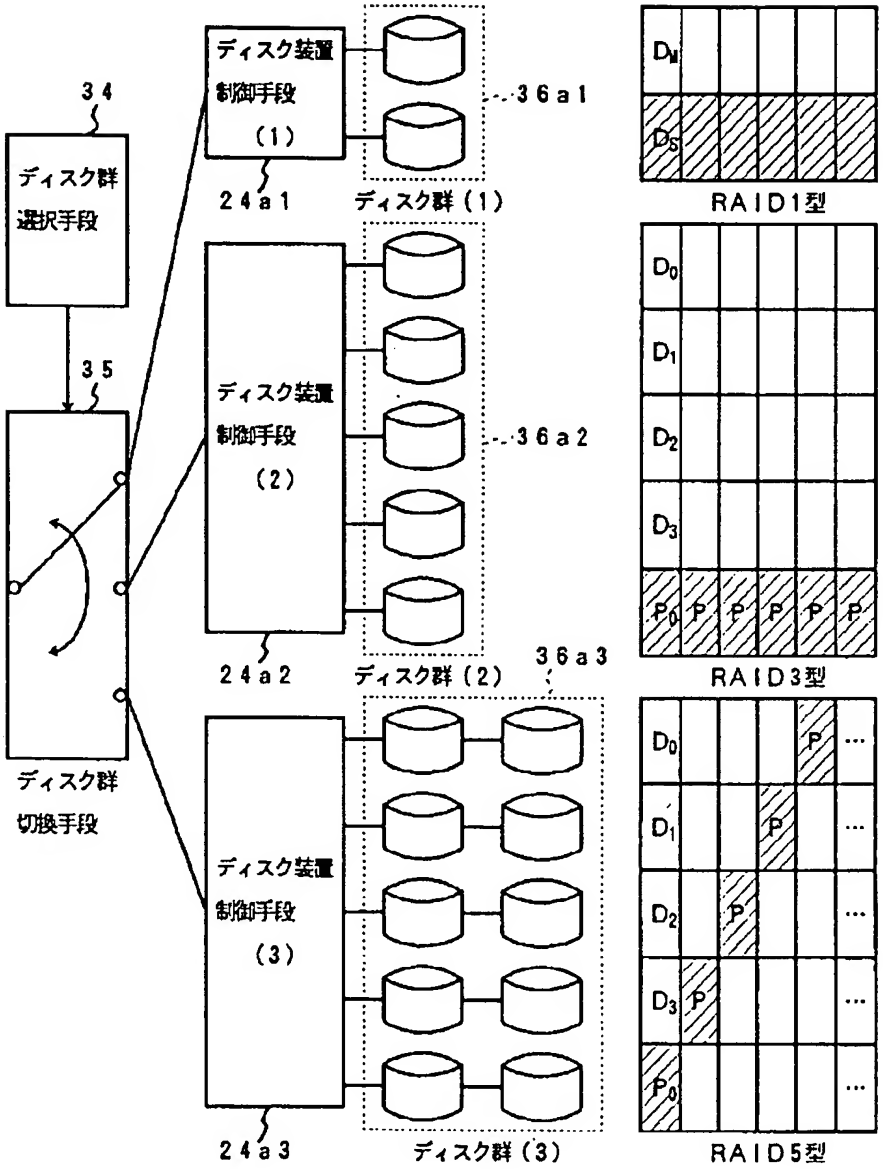
【図１８】

図18


上位計算機へ
２２
インタフェース制御手段
新規ファイル通知手段
２８
１８００
２３
２６
ディスク群
更新
制御手段
コマンド管理手段
新規ファイル
アドレス決定
手段
ファイル管理
テーブル
２７
３５
ディスク
装置群
切換手段
ディスク
装置群
選択手段
３４
ディスク管理
テーブル
（１）
２５ｂ１
ディスク管理
テーブル
（２）
２５ｂ２
ディスク管理
テーブル
（３）
２５ｂ３
ディスク装置
制御手段
（１）
ディスク装置
制御手段
（２）
ディスク装置
制御手段
（３）
２４ａ１
２４ａ２
２４ａ３
ディスク装置群
（１）
ディスク装置群
（２）
ディスク装置群
（３）
３６ａ１
３６ａ２
３６ａ３

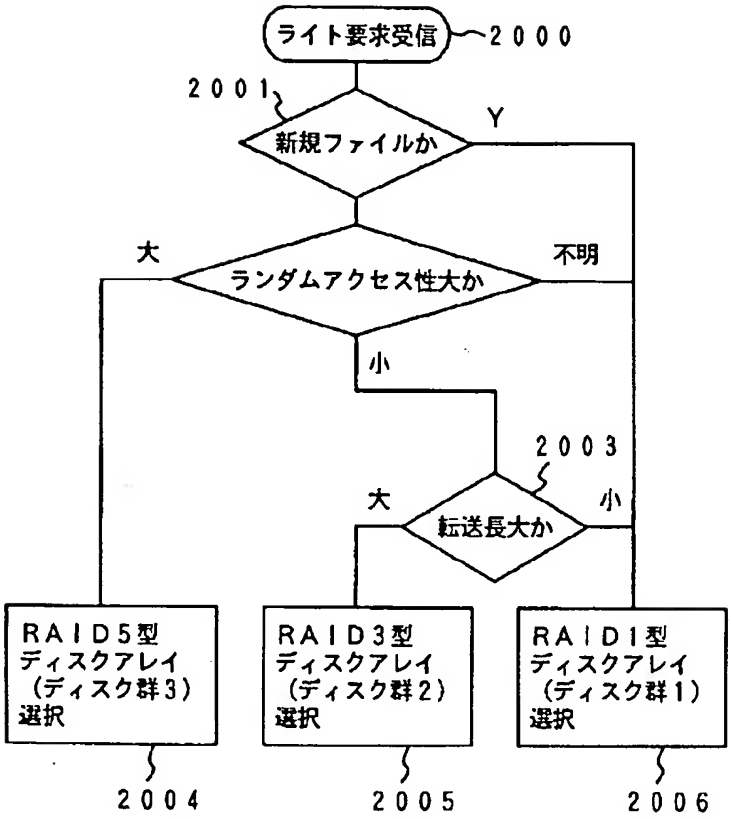
【図19】

図19


34
ディスク群
選択手段
35
ディスク群
切換手段
ディスク装置
制御手段
(1)
24a1
36a1
ディスク群(1)
ディスク装置
制御手段
(2)
24a2
36a2
ディスク群(2)
36a3
ディスク装置
制御手段
(3)
24a3
ディスク群(3)
RAID1型
RAID3型
RAID5型

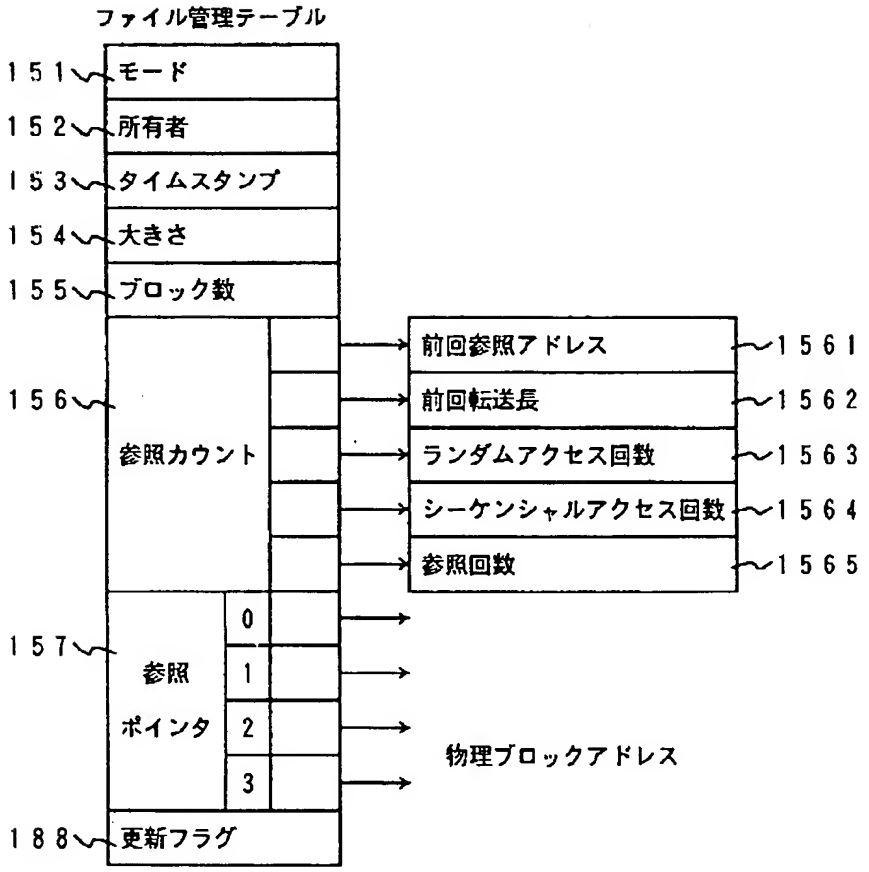
【図２０】

図20


ライト要求受信
2000
2001
新規ファイルか
Y
大
ランダムアクセス性大か
不明
小
2003
大
転送長大か
小
RAID5型
ディスクアレイ
（ディスク群3）
選択
RAID3型
ディスクアレイ
（ディスク群2）
選択
RAID1型
ディスクアレイ
（ディスク群1）
選択
2004
2005
2006

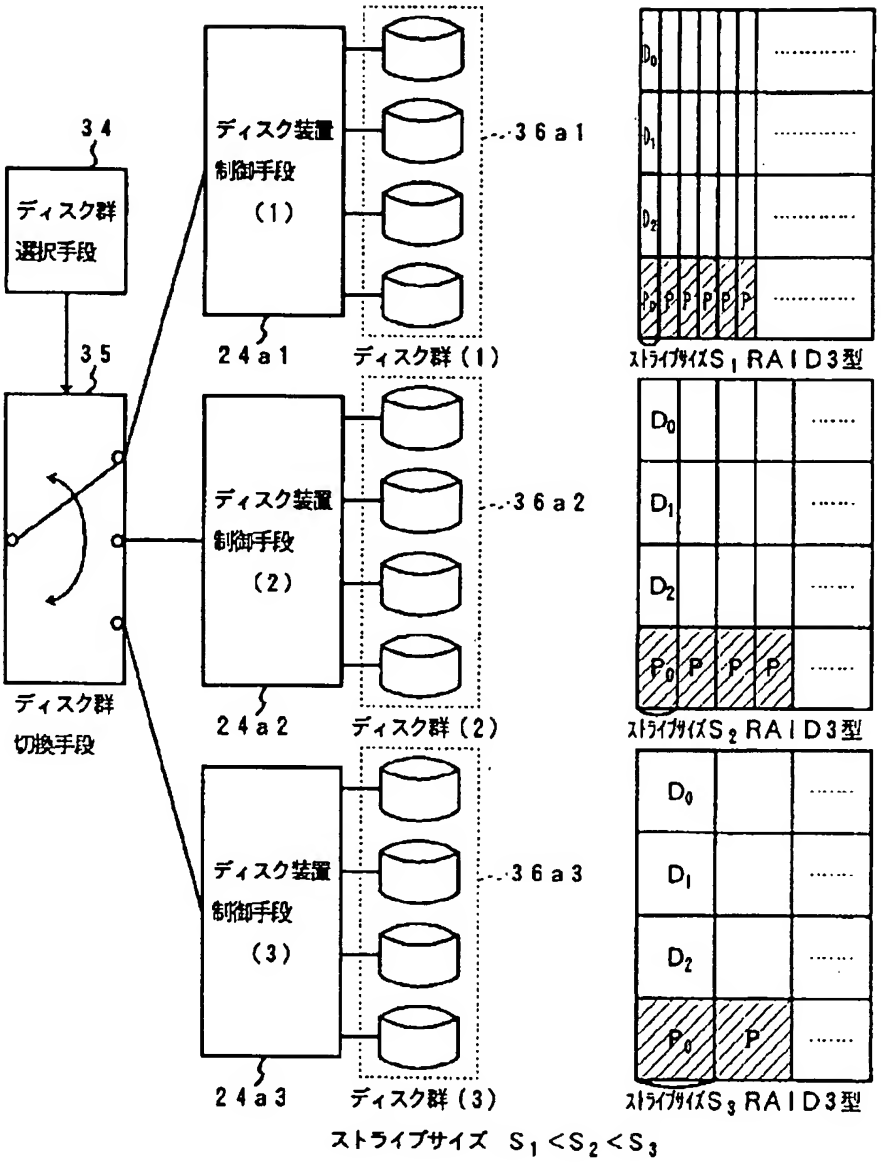
【図２１】

図21


ファイル管理テーブル
１５１
モード
１５２
所有者
１５３
タイムスタンプ
１５４
大きさ
１５５
ブロック数
１５６
参照カウント
前回参照アドレス
１５６１
前回転送長
１５６２
ランダムアクセス回数
１５６３
シーケンシャルアクセス回数
１５６４
参照回数
１５６５
１５７
参照
ポインタ
0
1
2
3
物理ブロックアドレス
１８８
更新フラグ

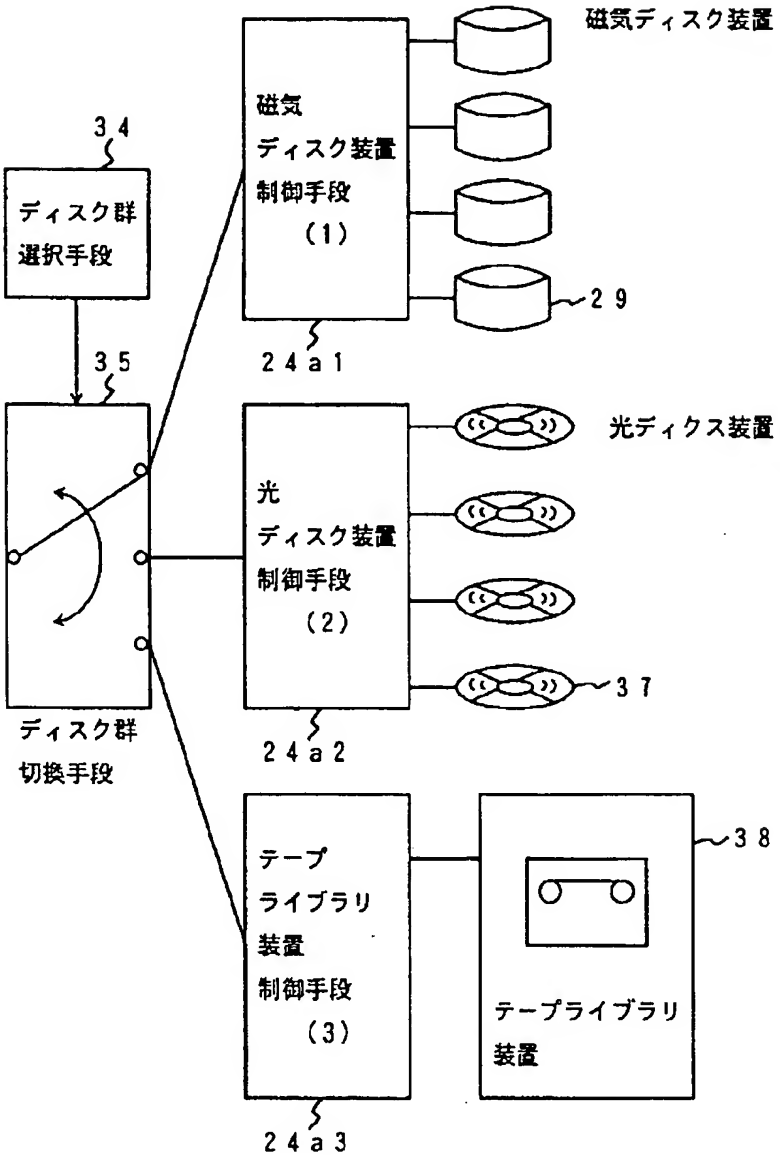
【図22】

図22


34
ディスク群
選択手段
35
ディスク群
切換手段
ディスク装置
制御手段
(1)
24a1
36a1
ディスク群(1)
ディスク装置
制御手段
(2)
24a2
36a2
ディスク群(2)
ディスク装置
制御手段
(3)
24a3
36a3
ディスク群(3)
D0
D1
D2
P0
P
ストライプサイズS1 RAID3型
ストライプサイズS2 RAID3型
ストライプサイズS3 RAID3型
ストライプサイズ S1<S2<S3

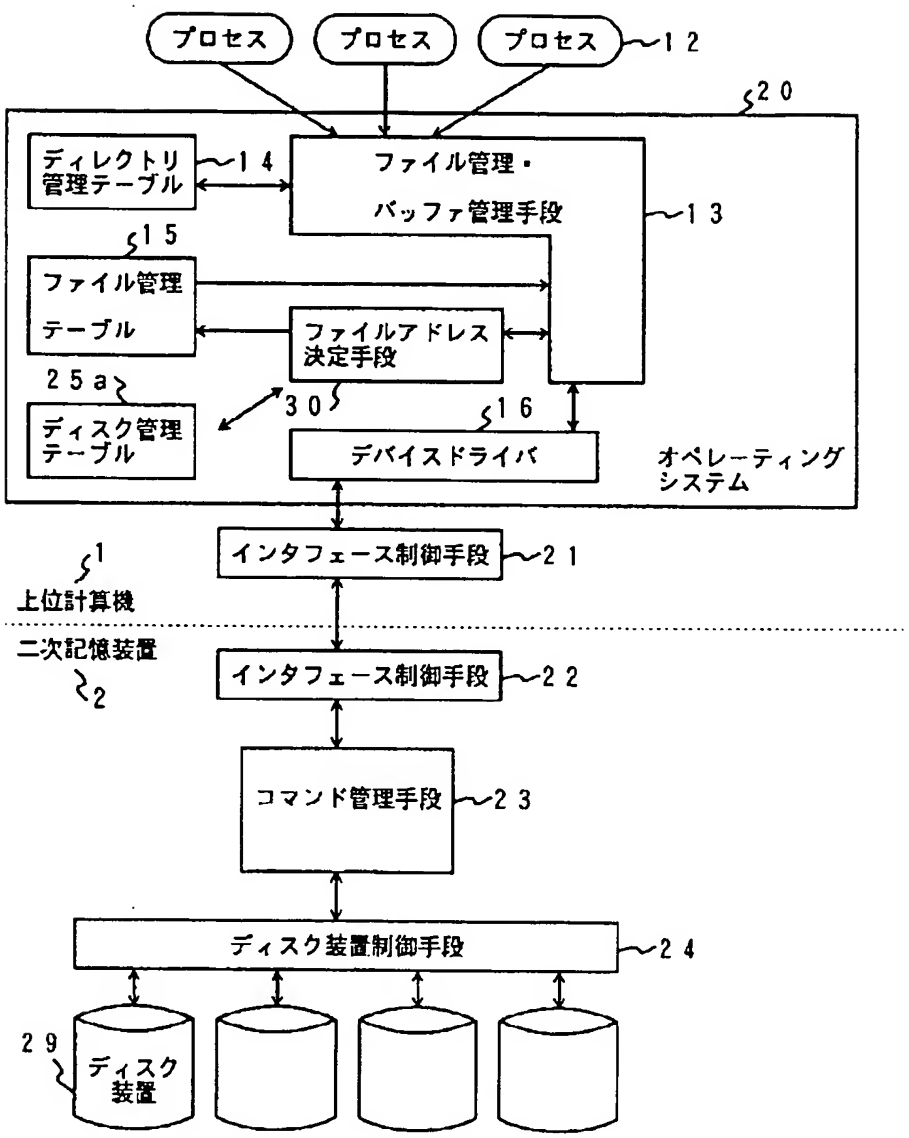
【図２３】

図23


磁気ディスク装置
磁気
ディスク装置
制御手段
（1）
２９
２４ａ１
３４
ディスク群
選択手段
３５
光ディスク装置
光
ディスク装置
制御手段
（2）
３７
２４ａ２
ディスク群
切換手段
テープ
ライブラリ
装置
制御手段
（3）
３８
テープライブラリ
装置
２４ａ３

【図24】

図24



フロントページの続き

(72)発明者　▲吉▼田　稔
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内